MANUAL PRÁTICO DE ANONIMIZAÇÃO DE DADOS DE PESQUISA COM O R

Autores
Marcelo Ribeiro-Alves
Carolina Mendes Franco

2022

Ministério da Saúde
FIOCRUZ
Fundação Oswaldo Cruz

Dados Internacionais de Catalogação na Publicação (CIP)
(Câmara Brasileira do Livro, SP, Brasil)

Ribeiro-Alves, Marcelo
Manual prático de anonimização de dados de pesquisa com o R [livro eletrônico] / Marcelo Ribeiro-Alves, Carolina Mendes Franco. -- Rio de Janeiro : Fundação Oswaldo Cruz, 2022.
PDF

Bibliografia.
ISBN 978-65-00-59103-3

1. Anonimização - Dados pessoais 2. Comunicação científica e tecnológica 3. Dados - Análise 4. Dados - Proteção 5. Divulgação de informação científica 6. Estatística - Métodos 7. Pesquisa científica 8. Proteção de dados - Direito - Brasil I. Franco, Carolina Mendes. II. Título.

22-139763 CDD-342.721

Índices para catálogo sistemático:

1. Anonimização : Proteção de dados pessoais : Direito 342.721

Henrique Ribeiro Soares - Bibliotecário - CRB-8/9314

Sumário

Manual Prático de Anonimização de Dados de Pesquisa com o R 5
I – Apresentação 5
II - Introdução 6
Seção 1 - Conceitos 7
1.1 – Classificação de Variáveis 7
1.2 – O que é divulgação? 9
1.3 – Observações sobre os métodos SDC 10
1.3.1 – Métodos para medir o risco de divulgação 10
1.3.1.1 – Risco individual 11
1.3.1.1.1 – Medidas de risco para variáveis-chave categóricas 11
1.3.1.1.1.1 – k-anonimato 12
1.3.1.1.1.2 – l-diversidade 13
1.3.1.1.1.3 – Algoritmo de Detecção de Exclusividade Especial (SUDA) 14
1.3.1.1.2 – Medidas de risco para variáveis-chave numéricas contínuas 15
1.3.1.1.2.1 – Vinculação de registros 15
1.3.1.1.2.2 – Medida de intervalo 16
1.3.1.1.2.3 – Detecção de valores extremos 17
1.3.1.2 – Risco Global 17
1.3.1.2.1 – Média das medidas de risco individuais 18
1.3.1.2.2 – Contagem de indivíduos com riscos maiores do que um determinado limiar 18
1.3.1.3 – Risco domiciliar 18
1.3.2 – Métodos de desidentificação de microdados 19
1.3.2.1 – Recodificação 19
1.3.2.2 – Supressão local 20
1.3.2.3 – Pós-aleatorização (PRAM) 20
1.3.2.4 – Microagregação 21
1.3.2.5 – Adição de Ruído 22
1.3.2.6 – Embaralhamento 23
1.3.3 – Métodos de medida da perda de informação 23
1.3.3.1 – Ferramentas gerais 24
1.3.3.2. Ferramentas específicas 24
1.4 – Risco em relação à utilidade dos dados e perda de informações 25
1.5 – O fluxo de trabalho 27
1.5.1 – Como determinar as variáveis-chave 30

1.5.2 – O Nível de Risco de Divulgação *versus* Perda de Informação 30
1.5.3 – Quais métodos SDC devem ser utilizados? 31
1.5.3.1 - Diretrizes Gerais para comparação de métodos 31
Seção 2 – Instalação do *software* R e do pacote 'sdcMicro' e outros pacotes 33
2.1. Passo a passo com o uso do R 33
2.2 – Funções de leitura em R 34
2.3 – Valores Ausentes 35
2.4 – Classes no R 35
2.5 – Objetos da classe sdcMicroObj 36
Tabela 1 - Nomes de *slots* e descrição desses *slots* do objeto sdcMicroObj 37
2.6 – Estrutura hierárquica 38
2.7 – Tempo de computação 39
Seção 3 – Aplicação de métodos SDC na desidentificação de microdados sintéticos 39
3.1 – Etapa 1: Necessidade de controle de divulgação 40
3.2 – Etapa 2: Preparação dos dados e exploração das características dos dados 40
Tabela 2 - Descrição das variáveis do subconjunto sintético de dados do estudo de caso. 44
3.3 – Etapa 3: Tipo de liberação 49
3.4 – Etapa 4: Cenários de intrusos e escolha das variáveis-chave 50
3.5 – Etapa 5: Utilização da chave de dados e seleção de medidas de utilidade 51
3.6 – Etapa 6a: Avaliar o risco de divulgação (nível domiciliar) 55
3.7 – Etapa 7a: Avaliação das medidas de utilidade (nível doméstico) 61
3.8 – Etapa 8a: Escolha e aplicação dos métodos SDC (variáveis domiciliares) 61
3.9 – Etapa 9a: Remensuração do risco 72
3.10 – Etapa 10a: Remensuração da utilidade 74
3.11 – Etapa 6b: Avaliação do risco de divulgação (nível individual) 77
3.12 – Etapa 7b: Avaliação das medidas de utilidade (nível individual) 79
3.13 – Etapa 8b: Escolha e aplicação dos métodos SDC (nível individual) 79
3.14 – Etapa 9b: Remensuração do Risco (nível individual) 84
3.15 – Etapa 10b: Remensuração da Utilidade (nível individual) 85
3.16 – Etapa 11: Auditoria e relatórios 85
3.17 – Etapa 12: Liberação de dados 86
Referências 88

Manual Prático de Anonimização de Dados de Pesquisa com o R

I – Apresentação

Este manual é uma introdução à desidentificação de informação pessoal em microdados de pesquisa usando ferramentas conjuntamente conhecidas como de controle de divulgação estatística, SDC, (*Statistical Disclosure Control)* – em inglês.

Para melhor compreensão, ele foi organizado em 3 seções, divididos da seguinte forma. Na seção 1 serão apresentados conceitos básicos e como pode ser elaborado um planejamento em um fluxo de trabalho geral. Na seção 2, há o passo a passo para a instalação do *software estatístico* aberto "R" e do pacote 'sdcMicro', que podem ser usados na aplicação de métodos SDC na desidentificação de microdados sintéticos, apresentados na seção 3.[1]

A utilização do programa R e do pacote 'sdcMicro' pressupõe o conhecimento básico de algumas expressões em inglês, tendo em vista que a linguagem da programação, seus comandos, ocorre neste idioma. Por isso, funções computacionais do pacote 'sdcMicro', e outras do programa R, são apresentadas no texto seguindo a notação nomeDaFunção, seguida por '()', e.g. freqCalc().

Este manual é produto da pesquisa denominada "Compatibilização ético-legal de pesquisas com dados pessoais e corporais", financiada pelo Programa Inova Fiocruz, Geração de Conhecimento – Novos Talentos.

Palavras-chave: Anonimização de dados, Estatística, Comunicação e Divulgação Científica,

Lei de Proteção de Dados de Caráter Pessoal, Privacidade, Saúde Pública.

[1] Uma abordagem mais aprofundada dos assuntos aqui tratados pode ser encontrada em "Templ, Matthias (2017). Statistical Disclosure Control for Microdata: Methods and Applications in R. Springer. ISBN 978-3-319-50272-4".

II - Introdução

A era em que vivemos, dos grandes bancos de dados, denominada em inglês como *Big Data,* com o rápido armazenamento e processamento de informações, vem trazendo melhorias em diversos aspectos, mas também gera impactos negativos.

Na saúde, as informações coletadas por anos e utilizadas em estudos epidemiológicos, podem aperfeiçoar diagnósticos, mas podem, também, trazer riscos de identificação, estigmatização e prejuízos aos indivíduos, se não forem adotadas cautelas que impeçam vazamento de dados.

A entrada em vigor da lei de proteção brasileira, nº 13.709/2018, estabeleceu parâmetros importantes para a proteção dos dados, no Brasil. Nela, há vários dispositivos que incentivam a anonimização dos dados pessoais, como, por exemplo, no artigo 7º, IV, nos estudos realizados por órgãos de pesquisa, "sempre que possível", inclusive, quando for indispensável, sem o consentimento dos titulares (artigo, 11, II, "c").

A lei brasileira citada também esclarece que se a anonimização dos dados for irreversível, estes dados nem serão considerados dados pessoais para a finalidade da lei, como se interpreta da leitura do artigo 12, lei nº 13.709/2018.

A não incidência da lei brasileira para os dados anonimizados de forma irreversível enuncia a importância da diferenciação feita pela lei entre as outras duas hipóteses. Portanto, quando a anonimização efetivada puder ser reversível, com meios exclusivamente próprios ou com esforços razoáveis (artigo 12, *caput)* ou quando for efetuada uma pseudonimização – isto é, o tratamento em que o dado é desconectado do indivíduo, através do uso de informação adicional mantida em separado em outro ambiente, controlado e seguro (artigo 13 §4) – continua incidindo a lei.

Este manual, portanto, busca auxiliar o pesquisador que pretende realizar a anonimização dos dados coletados com a finalidade de pesquisa e que pretende divulgar seu trabalho cientificamente e disponibilizá-los para terceiros, caso exigido.

Destacamos que existem várias maneiras de realizar este trabalho de "anonimização" e que a situação tecnológica que hoje pode permitir este trabalho, amanhã pode estar superada. A

própria lei realiza tais reflexões no §1º do artigo 12 e incisos III e XI do artigo 5º.

Observamos, ainda, que a exigência da disponibilização dos dados da pesquisa pode advir de um periódico ou ainda da própria instituição onde o pesquisador trabalha, o que lhe impõe a organização da sua pesquisa, desde o início, com a preocupação legal e ética de atender aos parâmetros da anonimização, levando em consideração a divulgação da pesquisa, ao final, sem prejuízo aos titulares dos dados pessoais.

Seção 1 - Conceitos

Um arquivo de microdados é um conjunto de dados que contém informações coletadas sobre unidades estatísticas. Exemplos de unidades incluem pessoas, domicílios ou empresas. Para cada unidade, um conjunto de variáveis é coletado e está disponível no conjunto de dados. Esta seção discute conceitos relacionados à divulgação de dados pessoais, métodos SDC e fornece um fluxo de trabalho que mostra como aplicar os métodos SDC aos microdados.

1.1 – Classificação de Variáveis

De acordo com os riscos de divulgação, as variáveis podem ser classificadas em três grupos, que não são necessariamente disjuntivos:

A – Os identificadores diretos são variáveis que identificam precisamente as unidades estatísticas. Por exemplo, os números da carteira de identificação civil, CPF, título eleitoral ou telefones fixos ou celular, nomes de pessoas e endereços são exemplos de identificadores diretos.

B – As variáveis-chave ou quase identificadores (*quasi-identifiers*) é um conjunto de variáveis que podem ser usadas para identificar unidades estatísticas. Por exemplo, pode ser possível identificar indivíduos usando uma combinação de variáveis tais como sexo, idade, região e ocupação. Outros exemplos de variáveis-chave são renda, estado de saúde, nacionalidade

ou preferências políticas. As variáveis-chave também são chamadas de identificadores implícitos ou quase-identificadores. Ao discutir os métodos SDC, é preferível distinguir entre variáveis-chave categóricas, contínuas e semicontínuas, com base na escala dessas variáveis.

Esta classificação é importante para determinar os métodos SDC apropriados para aquela variável, bem como a validade das medidas de risco.

B.1 – As variáveis-chave categóricas tomam valores sobre um conjunto finito, e quaisquer operações aritméticas que as utilizam geralmente não são significativas ou não são permitidas. Exemplos de variáveis categóricas são gênero, região e nível de instrução;

B.2 -- As variáveis-chave contínuas podem assumir um número infinito de valores em um determinado conjunto. Exemplos são renda, altura do corpo e tamanho do terreno. Variáveis contínuas podem ser transformadas em variáveis categóricas através da construção de intervalos (tais como faixas de renda);

B.3 – As variáveis-chave semicontínuas são variáveis contínuas que assumem valores que se limitam a um conjunto finito. Um exemplo é a idade medida em anos, que poderia assumir valores no conjunto {0, 1, ..., 100}. A natureza finita dos valores destas variáveis significa que elas podem ser tratadas como variáveis categóricas para fins de SDC.

C – Variáveis não identificadoras são variáveis que não se enquadram no conceito de identificadores diretos nem no conceito de variáveis-chave.

Além destas classificações de variáveis, o processo SDC classifica ainda mais as variáveis de acordo com sua sensibilidade ou confidencialidade. Tanto as variáveis-chave quanto as não identificadoras podem ser classificadas como sensíveis (ou confidenciais) ou não sensíveis (ou não-confidenciais). Ressalte-se que esta distinção não é importante para os identificadores diretos, já que estes são removidos dos dados divulgados.

D – As variáveis sensíveis contêm informações confidenciais que não devem ser divulgadas sem tratamento adequado. Para reduzir o risco na divulgação, pode-se recorrer a métodos de anonimização, tais como o SDC . Exemplos de variáveis sensíveis são renda, religião,

filiação política e saúde. É preciso destacar que uma variável pode ser sensível dependendo do contexto e do país.

E – As variáveis não sensíveis contêm informações não confidenciais sobre uma pessoa, tais como local de residência ou residência rural/urbana. A classificação de uma variável como não sensível, entretanto, não significa que ela não precisa ser considerada no processo SDC. Variáveis não sensíveis ainda podem servir como quase-identificadores quando combinadas com outras variáveis ou outros dados externos.

1.2 – O que é divulgação?

Em geral, a divulgação ocorre quando um intruso usa os dados divulgados para revelar informações previamente desconhecidas sobre um participante de pesquisa. Há três tipos diferentes de divulgação:

A - Divulgação de identidade: Neste caso, o intruso associa um indivíduo com um registro de dados compartilhados que contém informações sensíveis, ou seja, realiza a ligação com dados disponíveis externos.

A divulgação da identidade é possível através de identificadores diretos, combinações raras de valores nas variáveis-chave e conhecimento exato dos valores contínuos das variáveis-chave em bancos de dados externos. Para estes últimos, valores extremos de dados (por exemplo, valores extremamente altos de idade para um participante de pesquisa) levam a altos riscos de re-identificação.

B – Divulgação de características: Neste caso, o intruso é capaz de determinar algumas características de um indivíduo baseado nas informações disponíveis nos dados compartilhados. Por exemplo, se todos os participantes de 56 a 60 anos que se autodeclararam negros e eram residentes na região xxx estavam desempregados, então o intruso pode determinar o valor da variável “condição de trabalho”.

C – Divulgação inferencial: Neste caso, o intruso, embora com alguma incerteza, pode prever com alguma precisão o valor de uma ou mais características de um participante de pesquisa com os dados compartilhados.

Observamos que se a ligação baseada em uma série de identificadores for bem sucedida, o intruso terá acesso a todas as informações relacionadas a um conjunto de participantes dos dados compartilhados. Isto significa que um subconjunto de variáveis críticas pode ser explorado para divulgar tudo sobre um conjunto de participantes.

1.3 – Observações sobre os métodos SDC

Em geral, os métodos SDC utilizam técnicas emprestadas de outros campos. Os métodos SDC podem ser divididos em três grandes áreas:

(1) Métodos para medir o risco de divulgação;

(2) Métodos para tornar os microdados anônimos ou desidentificá-los; e,

(3) Métodos de comparação entre os dados originais e os modificados (métodos que medem a perda de informação).

1.3.1 – Métodos para medir o risco de divulgação

A avaliação do risco de divulgação é realizada com relação às fontes de dados disponíveis no ambiente onde o conjunto de dados deve ser divulgado. Neste cenário, o risco de divulgação é a possibilidade de reidentificar corretamente um indivíduo no arquivo de microdados liberado, combinando seus dados com um arquivo externo baseado em um conjunto de quase-identificadores.

A avaliação de risco é feita pela identificação dos chamados cenários de divulgação ou intrusão. Um cenário de divulgação descreve as informações potencialmente disponíveis para o invasor (por exemplo, dados do censo, listas eleitorais, registros populacionais ou dados coletados por empresas privadas) para identificar os respondentes e as formas como tais informações podem ser combinadas com o conjunto de microdados a serem liberados e usados para reidentificação de registros no conjunto de dados. Tipicamente, estes conjuntos de dados externos incluem identificadores diretos. Nesse caso, a reidentificação de registros

no conjunto de dados liberado leva à identidade e, possivelmente, à divulgação de atributos. O principal resultado da avaliação dos cenários de divulgação é a identificação de um conjunto de quase-identificadores (ou seja, variáveis-chave) que precisam ser tratados durante o processo SDC.

Um exemplo de um cenário de divulgação poderia ser o reconhecimento espontâneo de um respondente por um pesquisador. Por exemplo, ao passar pelos dados, o pesquisador reconhece uma pessoa com uma combinação incomum das variáveis “idade” e “estado civil”. É claro que isto só pode acontecer se a pessoa for bem conhecida ou for conhecida pelo pesquisador.

A avaliação do risco de divulgação é baseada nos quase-identificadores, que são identificados na análise dos cenários de risco de divulgação. O risco de divulgação depende diretamente da inclusão ou exclusão de variáveis no conjunto de quase-identificadores escolhidos. Esta etapa do processo SDC (fazer a escolha dos quase-identificadores) deve, portanto, ser abordada com muita reflexão e cuidado.

O primeiro passo do processo SDC, portanto, é realizar um exercício no qual se faz um inventário de todos os conjuntos de dados disponíveis no país. Os conjuntos de dados bem como as variáveis incluídas nestes conjuntos de dados são analisados. São estas informações que servirão como uma métrica chave ao decidir quais variáveis escolher como identificadores potenciais, assim como ditar o nível de SDC e os métodos necessários.

1.3.1.1 – Risco individual

1.3.1.1.1 – Medidas de risco para variáveis-chave categóricas

O principal foco da medição de risco para quase-identificadores categóricos é a divulgação de identidade. A medição do risco de divulgação é baseada na avaliação da probabilidade de reidentificação correta dos indivíduos nos dados divulgados. Usam-se medidas baseadas nos microdados reais a serem liberados. Em geral, quanto mais rara for uma combinação de valores dos quase-identificadores (ou seja, chave) de uma observação na amostra, maior será o risco de divulgação de identidade. Se um indivíduo tem uma combinação única de

valores de quase-identificadores e é chamado de "amostra única" e um intruso tente combinar um indivíduo que tenha uma chave relativamente rara dentro dos dados da amostra com um conjunto de dados externo no qual a mesma chave exista, terá maior probabilidade de encontrar uma combinação correta do que quando um número maior de indivíduos compartilha a mesma chave, pois indivíduos com as mesmas chaves têm a mesma frequência de amostragem.

Na prática, a abordagem de cálculo de frequências de chaves leva a estimativas de risco conservadoras, pois não leva em conta adequadamente os métodos de amostragem. Neste caso, as estimativas de risco de reidentificação podem ser muito estimadas. Se este risco superestimado for usado, os dados podem ser superprotegidos (ou seja, a perda de informação será maior do que o necessário) ao aplicar as medidas SDC.

1.3.1.1.1.1 – k-anonimato

A medida de risco k-anonimato (do inglês *k*-anonimity) se baseia no princípio de que, em um conjunto de dados seguro, o número de indivíduos que compartilham a mesma combinação de valores (chaves) de quase-identificadores categóricos devem ser maiores que um limite especificado k.

k-anonimato é uma medida de risco baseada nos microdados a serem liberados, uma vez que leva em conta apenas a amostra. Um indivíduo viola o k-anonimato se a contagem de frequência da amostra para a chave k for menor do que o limite especificado k. Por exemplo, se um indivíduo tem a mesma combinação de quase-identificadores que apenas outros dois indivíduos na amostra, estes indivíduos satisfazem o 3-anonimato, mas não alcançariam o 4 (ou >4)-anonimato, isto é, violariam o 4 (ou >4)-anonimato.

A medida de risco é o número de observações que violam o k-anonimato para um determinado valor de k. A medida de risco k-anonimato não considera os pesos de amostra, mas é importante considerar os pesos de amostra ao determinar o nível requerido de k-anonimato. Se os pesos de amostra forem grandes, um indivíduo no conjunto de dados representa mais indivíduos na população alvo, a probabilidade de uma correspondência

correta é menor e, portanto, o limiar exigido pode ser menor. Pesos de amostra grandes acompanham conjuntos de dados menores. Em um conjunto de dados menor, a probabilidade de encontrar outro registro com a mesma chave é menor do que em um conjunto de dados maior. Esta probabilidade está relacionada ao número de registros na população com uma determinada chave através dos pesos de amostra.

Se um conjunto de dados satisfizer o k-anonimato, um intruso sempre encontrará pelo menos k indivíduos com a mesma combinação de quase-identificador. k-anonimato é frequentemente um requisito necessário para a desidentificação de um conjunto de dados antes da liberação, mas não é necessariamente um requisito suficiente. A medida de k-anonimato é apenas baseada em contagens de frequência e não leva em conta (diferenças em) pesos de amostra. Muitas vezes o k-anonimato é alcançado aplicando primeiro a recodificação e depois a supressão local, e em alguns casos por microagregação, antes de usar outras medidas de risco e métodos de divulgação para reduzir ainda mais o risco de divulgação.

1.3.1.1.1.2 – l-diversidade

O k-anonimato tem sido criticado por não ser suficientemente restritivo, uma vez que informações sensíveis podem ser divulgadas mesmo que os dados satisfaçam o k-anonimato. Isto pode ocorrer nos casos em que os dados contenham variáveis categóricas sensíveis (não identificadoras) que tenham o mesmo valor para todos os indivíduos que compartilham a mesma chave. Exemplos de tais variáveis sensíveis são aqueles que contêm informações sobre o estado de saúde de um indivíduo.

O conceito de l-diversidade - do inglês l-diversity - trata desta deficiência do k-anonimato. Então, um conjunto de dados satisfaz a l-diversidade se para cada chave k houver pelo menos l valores diferentes para cada uma das variáveis sensíveis. O nível requerido de l-diversidade depende do número de valores possíveis que a variável sensível pode tomar. Se a variável sensível for uma variável binária, o nível mais alto de l-diversidade que pode ser alcançado é 2.

Uma amostra única sempre atenderá apenas 1-diversidade. Portanto, l-diversidade é útil se os dados contiverem variáveis sensíveis categóricas que não sejam quase-identificadores em si. Não é possível selecionar quase-identificadores para calcular a l-diversidade. A l-diversidade tem que ser calculada para cada variável sensível separadamente.

1.3.1.1.1.3 – Algoritmo de Detecção de Exclusividade Especial (SUDA)

As medidas de risco discutidas anteriormente dependem da identificação de variáveis-chave para as quais pode haver informações disponíveis de outras fontes ou outros conjuntos de dados, que, quando combinadas com os dados atuais, podem levar a uma reidentificação. Na prática, porém, pode nem sempre ser possível realizar um inventário de todos os conjuntos de dados disponíveis e suas variáveis para avaliação de todas as ligações e riscos externos conhecidos.

Para superar isto, uma medida heurística alternativa, baseada em unidades especiais, foi desenvolvida para determinar o grau de risco de um registro, o que leva a uma métrica ou pontuação SUDA - do inglês *Special Uniqueness Detection Algorithm*.

Os algoritmos SUDA são baseados no conceito de singularidade especial. O algoritmo SUDA identifica toda a amostra mínima única (MSU) - já que qualquer subconjunto menor deste conjunto de variáveis não é único na amostra - que são usadas para atribuir uma pontuação SUDA a cada registro. Esta pontuação indica o quão "arriscado" é um registro.

O risco potencial dos registros é determinado com base em duas observações: (1) Quanto menor o tamanho da MSU dentro de um registro (ou seja, quanto menos variáveis forem necessárias para atingir a singularidade de registros), maior o risco daquele registro; e, (2) Quanto maior o número de MSUs em um conjunto de dados, maior é o risco da identificação de registros nessa base.

Em vez de substituir as medidas de risco introduzidas nas subseções anteriores, a pontuação SUDA deve ser empregada como um método complementar, particularmente útil em situações em que é difícil fazer um inventário de todos os conjuntos de dados já disponíveis e suas variáveis.

1.3.1.1.2 – Medidas de risco para variáveis-chave numéricas contínuas

O princípio da raridade ou singularidade das combinações de quase-identificadores (chaves) não é útil para variáveis contínuas, pois é provável que todos ou muitos indivíduos tenham chaves únicas. Portanto, outras abordagens são exploradas para medir o risco de divulgação das variáveis contínuas. Estes métodos são baseados na singularidade dos valores na vizinhança dos valores originais.

A unicidade é definida de diferentes maneiras: em termos absolutos (medida de intervalo - do inglês *interval measure*) ou relativos (vinculação de registro - do inglês *record linkage*).

A maioria das medidas são obtidas *a posteriori*: elas são avaliadas após a desidentificação dos dados brutos, comparando os dados tratados com os dados brutos e avaliando, para cada indivíduo, a distância entre os valores nos dados brutos e os dados tratados. Isto significa que estes métodos não são úteis para identificar indivíduos em risco dentro dos dados brutos, mas mostram a distância/diferença entre o conjunto de dados antes e depois da desidentificação e podem, portanto, ser interpretados como avaliação do método de desidentificação.

Finalmente, as medidas de risco para quase-identificadores contínuos também se baseiam na detecção de valores extremos —

do inglês *outliers*. Os valores extremos têm um papel importante na reidentificação desses registros.

1.3.1.1.2.1 – Vinculação de registros

A vinculação de registros é um método que avalia *a posteriori* o número de vinculações corretas ao ligar os valores perturbados com os valores originais. O algoritmo de vinculação é baseado na distância entre os valores originais e os perturbados (ou seja, vinculação de registro baseada na distância). Os valores perturbados são combinados com o indivíduo mais próximo. É importante observar que este método não dá informações sobre o risco inicial, mas é antes uma medida para avaliar o algoritmo de perturbação, ou seja, é projetado

para indicar o nível de incerteza introduzido na variável, contando o número de registros que poderiam ser combinados corretamente.

Os algoritmos de vinculação de registros diferem em relação a qual medida de distância é usada. Quando uma variável tem escala muito diferente de outras variáveis contínuas no conjunto de dados, é recomendável redimensionar as variáveis antes de usar a vinculação entre registros. Escalas muito diferentes podem levar a resultados indesejados ao medir a distância multivariada entre registros com base em várias variáveis contínuas.

1.3.1.1.2.2 – Medida de intervalo

A aplicação bem sucedida de um método SDC deve resultar em valores perturbados que são considerados não muito próximos de seus valores iniciais; se o valor for relativamente próximo, a reidentificação pode ser relativamente fácil.

Na aplicação de medidas de intervalo são criados intervalos em torno de cada valor perturbado e então é feita uma determinação se o valor original dessa observação perturbada está contido nesse intervalo. Os valores que estão dentro do intervalo em torno do valor inicial após a perturbação são considerados muito próximos do valor inicial e, portanto, inseguros e precisam de mais perturbação. Os valores que estão fora dos intervalos são considerados seguros.

O tamanho do intervalo é baseado no desvio padrão das observações e em um parâmetro de escala. O tamanho dos intervalos é k vezes o desvio padrão. Quanto maior k, maiores são os intervalos e, portanto, maior o número de observações dentro do intervalo construído em torno de seus valores originais e maior é a medida de risco.

O resultado 1 indica que todas (100%) das observações estão fora do intervalo de 0,1 vezes o desvio padrão em torno dos valores originais. Para a maioria dos valores, esta é uma abordagem satisfatória. No entanto, não é uma medida suficiente para os valores extremos. Após a perturbação, os valores extremos permanecerão extremos e serão facilmente reidentificáveis, mesmo que estejam suficientemente longe de seus valores iniciais. Portanto, os valores extremos devem ser tratados com cautela.

1.3.1.1.2.3 – Detecção de valores extremos

Os valores extremos são importantes para medir o risco de reidentificação em microdados contínuos. Os dados contínuos são frequentemente enviesados, especialmente os enviesados para a direita. Isto significa que existem alguns valores extremos em relação às outras observações da mesma variável. Exemplos são a renda em dados domiciliares, onde apenas poucos indivíduos/famílias podem ter renda muito alta. Em casos como este, mesmo que estes valores sejam perturbados, ainda assim pode ser fácil identificar estes valores aberrantes, uma vez que eles permanecerão os maiores valores, mesmo após a perturbação. A perturbação pode até criar incerteza quanto ao valor exato, mas como o valor começou muito mais distante de outras observações, ainda assim pode ser fácil de vincular ao indivíduo de alta renda. Exemplo seria o único médico em uma área geográfica com alta renda. Portanto, a identificação de valores extremos em dados contínuos é um passo importante na identificação de indivíduos em alto risco. Na prática, a identificação dos valores de uma variável contínua que são maiores do que uma variável pré-determinada p%-percentil poderia ajudar a identificar valores extremos, e, portanto, unidades com maior risco de identificação. O valor de p depende do enviesamento dos dados.

Uma segunda abordagem para a detecção de valores extremos é uma medida *a posteriori* comparando os dados tratados e brutos. Um intervalo é construído em torno dos valores perturbados, como descrito na subseção anterior. Se os valores originais caírem no intervalo em torno dos valores perturbados, os valores perturbados são considerados inseguros por estarem muito próximos dos valores originais. Há diferentes maneiras de construir tais intervalos, tais como intervalos baseados em ordenação e intervalos baseados em desvios padrão.

1.3.1.2 – Risco Global

Para construir uma medida de risco agregada global para o conjunto de dados completo, podemos agregar as medidas de risco individuais de várias maneiras. As medidas de risco global devem ser usadas com cautela. Por trás de um risco global aceitável pode esconder

alguns registros de muito alto risco que são compensados por muitos registros de baixo risco.

1.3.1.2.1 – Média das medidas de risco individuais

Uma maneira simples de agregar as medidas de risco individuais é tomar a média de todos os indivíduos da amostra, que é igual à soma de todas as chaves da amostra - se multiplicada pelas frequências de amostragem dessas chaves - e dividida pelo tamanho da amostra.

1.3.1.2.2 – Contagem de indivíduos com riscos maiores do que um determinado limiar

Todos os indivíduos pertencentes à mesma chave têm o mesmo risco individual. Outra forma de expressar o risco total na amostra é o número total de observações que excedem um determinado limiar de risco individual. A definição do limiar pode ser absoluta (por exemplo, todos aqueles indivíduos que têm um risco de divulgação superior a 0,05 ou 5%) ou relativa (por exemplo, todos aqueles indivíduos com riscos superiores ao quartil superior de risco individual).

1.3.1.3 – Risco domiciliar

Em muitas pesquisas sociais, os dados têm uma estrutura hierárquica onde um indivíduo pertence a uma entidade de nível superior. Exemplos típicos são domicílios em pesquisas sociais ou alunos em escolas. A reidentificação de um membro da família pode levar à reidentificação dos outros membros da família, também. Portanto, é fácil ver que se levarmos em conta a estrutura familiar, o risco de reidentificação é o risco de que pelo menos um dos membros da família seja reidentificado.

O risco hierárquico ou familiar não pode ser inferior ao risco individual e o risco familiar é sempre o mesmo para todos os membros da família. O risco domiciliar deve ser usado nos

casos em que os dados contêm uma estrutura hierárquica, ou seja, quando uma estrutura domiciliar está presente nos dados.

1.3.2 – Métodos de desidentificação de microdados

1.3.2.1 – Recodificação

A recodificação global - do inglês *global recoding* - é um método não-perturbativo que pode ser aplicado tanto a variáveis chave categóricas como contínuas. A ideia básica da recodificação de uma variável categórica é combinar várias categorias em uma nova categoria, menos informativa. Se o método for aplicado a uma variável contínua, isso significa discretizar a variável.

Casos frequentes de seu uso são a divisão de uma variável contendo rendimentos em grupos de renda. O objetivo é reduzir o número total de resultados possíveis de uma variável. Tipicamente, a recodificação é aplicada a variáveis categóricas onde o número de categorias com poucas observações, ou seja, categorias extremas, - por exemplo, pessoas com mais de 100 anos - é reduzido.

Um caso especial de recodificação é a codificação superior e inferior, que pode ser aplicada a variáveis ordinais e categóricas. A ideia para esta abordagem é que todos os valores acima (ou seja, codificação superior) e/ou abaixo (ou seja, codificação inferior) de um valor limiar pré-especificado, sejam combinados em uma nova categoria. Um caso típico de uso para codificação superior é recodificar todos os valores de uma variável contendo idade em anos que estejam acima de 80 anos em uma nova categoria 80+.

A função globalRecode() pode ser aplicada na biblioteca sdcMicro para executar tanto a recodificação global quanto a codificação superior/inferior.

1.3.2.2 – Supressão local

A supressão local - do inglês *local suppression* - é um método não-perturbativo que normalmente é aplicado a variáveis categóricas para suprimir certos valores, em, pelo menos, uma delas. Normalmente, as variáveis de entrada fazem parte do conjunto de variáveis-chave que também é usado para o cálculo de riscos individuais. Os valores individuais são suprimidos de forma que o conjunto de variáveis com um padrão específico seja aumentado. A supressão local é frequentemente usada para alcançar o *k*-anonimato. Usando a função localSupp() da biblioteca sdcMicro, é possível suprimir os valores de uma variável-chave para todas as unidades com riscos individuais acima de um limiar pré-definido, dado um cenário de divulgação. Este procedimento requer a intervenção do usuário, definindo o limiar. Para suprimir automaticamente uma quantidade mínima de valores nas variáveis-chave para atingir o *k*-anonimato, pode-se usar a função localSuppression(). Este algoritmo também permite a especificação de um vetor de importância dependente do usuário que determina quais variáveis-chave são preferidas ao escolher os valores que precisam ser suprimidos. Nesta implementação, um algoritmo heurístico é chamado para suprimir o menor número possível de valores.

É possível especificar uma ordenação desejada de variáveis-chave em termos de importância, que o algoritmo leva em conta. É também possível especificar variáveis-chave que são consideradas de tal importância que quase nenhum valor para estas variáveis é suprimido.

1.3.2.3 – Pós-aleatorização (PRAM)

Pós-aleatorização ou PRAM - do inglês *post randomization* - é um método perturbador e probabilístico que pode ser aplicado a variáveis categóricas. A ideia é que os valores de uma variável categórica no arquivo de microdados original sejam transformados em outras categorias, levando em conta probabilidades de transição pré-definidas. Este processo é normalmente modelado usando uma matriz de transição conhecida.

Para cada categoria de uma variável categórica, esta matriz lista as probabilidades de mudança em outras categorias possíveis. Como exemplo, considere uma variável com

apenas 3 categorias: A1, A2 e A3. A transição de um valor da categoria A1 para a categoria A1 é, por exemplo, fixa com probabilidade p1 = 0,85, o que significa que somente com probabilidade p1 = 0,15, um valor de A1 pode ser alterado para A2 ou A3. A probabilidade de uma mudança da categoria A1 para A2 pode ser fixa com probabilidade p2 = 0,1 e muda de A1 para A3 com probabilidade p3 = 0,05. As probabilidades de mudar valores da classe A2 para outras classes e para A3, respectivamente, devem ser especificadas com antecedência. Todas as probabilidades de transição devem ser armazenadas em uma matriz que seja a principal entrada para a função pram() na biblioteca sdcMicro. A PRAM é aplicada a cada observação de forma independente e aleatória. Isto significa que são obtidas soluções diferentes para cada execução de PRAM se nenhuma semente for especificada para o gerador de números aleatórios.

Uma vantagem do procedimento da PRAM é a flexibilidade do método. Uma vez que a matriz de transição pode ser especificada livremente como parâmetro da função, todos os efeitos desejados podem ser modelados. Por exemplo, é possível proibir mudanças de uma categoria para outra, definindo a probabilidade correspondente na matriz de transição para 0. Na biblioteca sdcMicro há também a função pram_strat(), que permite que a PRAM seja executada independentemente em subgrupos do microconjunto de dados. Neste caso, o usuário precisa selecionar a variável de estratificação que define os subgrupos. Se a especificação desta variável for omitida, o procedimento PRAM é aplicado a todas as observações do conjunto de dados.

1.3.2.4 – Microagregação

A microagregação (do inglês *global microaggregation*) é um método perturbador que é tipicamente aplicado a variáveis contínuas. A ideia é que os registros sejam divididos em grupos. Dentro de cada grupo, os valores de cada variável são agregados. Os valores individuais dos registros para cada variável são substituídos pelo valor de agregação do grupo, que muitas vezes é a média, mas outros métodos robustos são possíveis. Dependendo do método escolhido, na função *microaggregation()* da biblioteca 'sdcMicro', parâmetros adicionais podem ser especificados. Por exemplo, é possível especificar o número de observações que devem ser agregadas, bem como a estatística usada para

calcular a agregação. Também é possível realizar a microagregação de forma independente para grupos pré-definidos ou usar métodos de cluster/aglomeração para se obter o agrupamento.

1.3.2.5 – Adição de Ruído

A adição de ruído - do inglês *adding noise* - é um método de proteção perturbativa para microdados, que normalmente é aplicado a variáveis contínuas. Esta abordagem protege os dados contra a correspondência exata com arquivos externos se, por exemplo, informações sobre variáveis específicas estiverem disponíveis a partir de registros. Embora esta abordagem pareça simples em princípio, muitos algoritmos diferentes podem ser usados para sobrepor os dados com o ruído estocástico. É possível adicionar ruído aleatório não correlacionado. Neste caso, o ruído é geralmente distribuído e a variação do termo 'ruído' é proporcional à variação do vetor de dados original.

Na adição de ruído não relacionado, as variâncias e coeficientes de correlação entre variáveis não são preservados. Esta propriedade estatística é respeitada, entretanto, se método(s) de ruído correlacionado(s) for(em) aplicado(s).

Para o método de ruído correlacionado, o termo ruído é derivado de uma distribuição com uma matriz de covariância que é proporcional à matriz de covariância dos microdados originais. No caso de adição de ruído correlacionado, os coeficientes de correlação são preservados e, pelo menos, a matriz de covariância pode ser consistentemente estimada a partir dos dados perturbados. A estrutura dos dados pode diferir muito, no entanto, se a suposição de normalidade for violada. Como este é praticamente sempre o caso quando se trabalha com conjuntos de dados do mundo real, uma versão robusta do método de ruído correlacionado está disponível na função addNoise() incluída na biblioteca 'sdcMicro'.

1.3.2.6 – Embaralhamento

O embaralhamento (do inglês *shuffling*) sintetiza valores das variáveis-chave contínuas condicionadas a variáveis independentes não confidenciais. Após a simulação dos novos valores para as variáveis-chave contínuas, é aplicado o mapeamento reverso. Isto significa que os valores classificados dos valores simulados são substituídos pelos valores classificados dos dados originais.

Para explicar este conceito teórico de forma mais prática, podemos assumir que temos duas variáveis contínuas contendo informações sensíveis sobre renda e poupança. Estas variáveis são utilizadas como variáveis dependentes em um modelo regressor onde variáveis adequadas são tomadas como preditores (independentes), como idade, ocupação, raça e escolaridade. É claro que é crucial encontrar um bom modelo com bom poder de previsão. Novos valores para as variáveis-chave contínuas, renda e economia, são preditos com base neste modelo. Entretanto, estes valores preditos esperados não são usados para substituir os valores originais, mas sim um embaralhamento dos valores originais usando os valores preditos como referência é realizado.

Esta abordagem de mapeamento reverso é aplicada a cada variável sensível. Vários métodos estão disponíveis para o embaralhamento na função shuffling() da biblioteca 'sdcMicro', e o argumento padrão (ds) é recomendado para uso.

1.3.3 – Métodos de medida da perda de informação

A utilidade da medição dos dados do conjunto de microdados após a aplicação dos métodos de limitação da divulgação é encorajada para avaliar o impacto desses métodos. Dados anonimizados devem ter quase a mesma estrutura dos dados originais e devem permitir qualquer análise com alta precisão.

1.3.3.1 – Ferramentas gerais

Para avaliar a precisão, utilizam-se várias estimativas clássicas, como médias e covariâncias. Usando a função dUtility() da biblioteca 'sdMicro', é possível calcular diferentes medidas com base nas distâncias clássicas ou robustas para variáveis em escala contínua. As estimativas são calculadas tanto para os dados originais como para os dados perturbados e depois comparados.

1.3.3.2. Ferramentas específicas

Na prática, não é possível criar um arquivo anônimo com a mesma estrutura que o arquivo original. Um objetivo importante, entretanto, deve ser sempre que a diferença nos resultados das estatísticas mais importantes baseadas em dados anonimizados e originais deve ser muito pequena ou até mesmo zero. Assim, o objetivo é medir a utilidade dos dados com base em indicadores de padrão, que é, em geral, uma abordagem melhor para avaliar a qualidade dos dados do que a aplicação de ferramentas gerais.

O primeiro passo na avaliação da qualidade é avaliar quais usuários dos dados subjacentes estão analisando e depois tentar determinar as estimativas mais importantes, ou indicadores de padrão. Deve ser dada ênfase especial aos indicadores de padrão que levam em conta as variáveis mais importantes do microconjunto de dados. Os indicadores que se referem às variáveis mais sensíveis dentro dos microdados também devem ser calculados. O procedimento geral é bastante simples e pode ser descrito nas seguintes etapas: (1) Seleção de um conjunto de indicadores de padrão; (2) Escolha de um conjunto de critérios sobre como comparar os indicadores; (3) Cálculo de todos os indicadores de padrão dos microdados originais; (4) Cálculo dos indicadores de padrão sobre o conjunto de microdados protegidos; (5) Comparação de propriedades estatísticas, tais como estimativas pontuais, desvios ou sobreposições em intervalos de confiança para cada indicador de *benchmarking*; e, (6) Avaliação se a utilidade dos dados do microconjunto de dados protegidos é suficientemente boa para ser usada pelos pesquisadores. Se a avaliação da qualidade na última etapa for satisfatória, o conjunto de microdados anonimizados está pronto para ser publicado. Se os desvios dos principais indicadores calculados a partir dos dados originais e

protegidos forem muito grandes, o procedimento de desidentificação deve ser reiniciado e modificado.

Normalmente, a avaliação é focada nas propriedades das variáveis numéricas, dados, microdados não modificados e modificados.

É claro que também é possível rever o impacto da supressão local ou recodificação que foi realizada para reduzir os riscos individuais de reidentificação.

Outra possibilidade de avaliar a utilidade dos dados das variáveis numéricas é definir um modelo que se ajuste aos microdados originais, não modificados. A ideia é prever variáveis importantes e sensíveis usando este modelo tanto para o conjunto de microdados original como para o modificado como um primeiro passo.

Em um segundo passo as propriedades estatísticas dos resultados do modelo, tais como as diferenças nas estimativas pontuais ou variâncias, são comparadas para as previsões entre os microdados originais e modificados, então a qualidade resultante é avaliada. Se os desvios forem suficientemente pequenos, pode-se passar à divulgação do conjunto de microdados seguros e protegidos. Caso contrário, devem ser feitos ajustes no procedimento de proteção.

Além disso, é interessante avaliar o conjunto de indicadores de padrão não apenas para todo o conjunto de dados, mas também independentemente para subconjuntos dos dados. Neste caso, os microdados são particionados em subconjuntos. A avaliação dos indicadores de padrão é então realizada para cada um dos subconjuntos e os resultados são avaliados através da revisão das diferenças entre indicadores para dados originais e dados modificados em cada subconjunto.

1.4 – Risco em relação à utilidade dos dados e perda de informações

O objetivo da SDC é compartilhar um conjunto de dados com alta utilidade de dados e seguro, ou seja, com baixo risco de se vincular informações pessoais aos participantes da

pesquisa. No mundo real, especialistas em desidentificação de dados devem basear suas decisões sobre risco e utilidade de dados considerando o seguinte:

A– Qual é a situação legal em relação à privacidade dos dados?

B– Qual é a sensibilidade das informações de dados e quem tem acesso ao arquivo de dados desidentificado?

Normalmente, as leis consideram dois tipos de usuários de dados: usuários de universidades e outras organizações de pesquisa, e usuários em geral, ou seja, o público. No primeiro caso, frequentemente são feitos contratos especiais entre usuários de dados e produtores de dados. Normalmente, estes contratos restringem o uso dos dados a fins muito específicos e permitem sua manipulação, somente dentro de ambientes de trabalho seguros. Para estes usuários, arquivos de microdados anônimos são chamados de arquivos de uso científico, enquanto os dados para o público são chamados de arquivos de uso público.

Naturalmente, o risco de divulgação de um arquivo de uso público precisa ser muito baixo, muito menor do que os riscos correspondentes em arquivos de uso científico. Para arquivos de uso científico, a utilidade dos dados é, geralmente, consideravelmente maior do que a utilidade dos dados para arquivos de uso público.

Outro aspecto que deve ser considerado é a sensibilidade do conjunto de dados. Os dados sobre tratamentos médicos individuais são mais sensíveis do que os valores de rotatividade e o número de funcionários de um hospital, por exemplo. Se os dados contiverem informações muito sensíveis, os microdados devem ter maior segurança do que os dados que contêm apenas informações que não são susceptíveis de serem atacadas por intrusos.

C– Qual método é adequado para qual finalidade? Os métodos de Controle de Divulgação Estatística implicam sempre em remover ou modificar variáveis selecionadas. A utilidade dos dados é reduzida em troca de mais proteção. Enquanto a aplicação de alguns métodos específicos resulta em baixo risco de divulgação e grande perda de informações, outros métodos podem fornecer dados com riscos aceitáveis e baixos de divulgação.

Recomendações gerais não podem ser dadas aqui, pois a solidez e fraqueza dos métodos depende do conjunto de dados utilizados. As decisões sobre quais variáveis serão

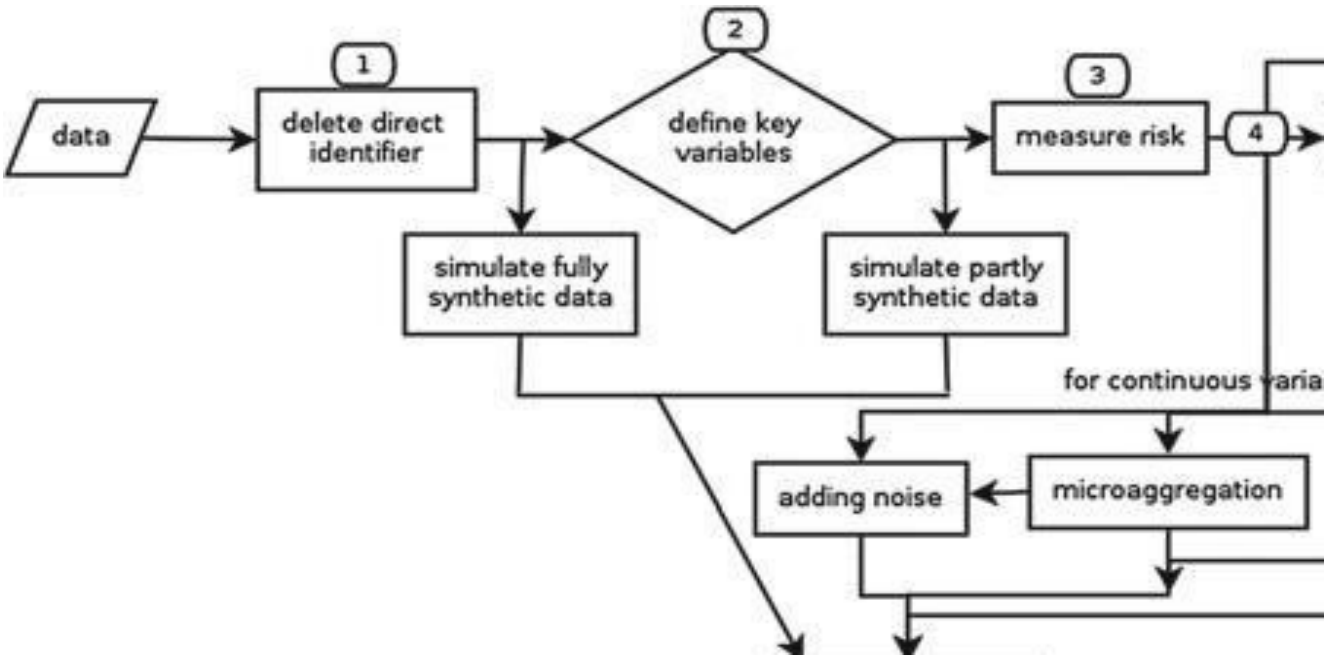
modificadas e qual método será utilizado resultam em certa medida de decisões discricionárias ou até arbitrárias, mas também de um conhecimento prévio do que os usuários farão com os dados. Geralmente, quando se tem apenas poucas variáveis-chave categóricas no conjunto de dados, recomenda-se a recodificação e supressão local para se obter baixo risco de divulgação em variáveis-chave categóricas. Além disso, no caso de variáveis-chave em escala contínua, a microagregação é fácil de se aplicar e de compreender e dá bons resultados. Para usuários mais experientes, o embaralhamento pode muitas vezes dar os melhores resultados, uma vez que há uma forte relação entre as variáveis-chave e outras variáveis do conjunto de dados. No caso de muitas variáveis-chave categóricas, a pós-aleatorização ou PRAM pode ser aplicada a várias dessas variáveis.

Métodos como a PRAM podem fornecer alto ou baixo risco de divulgação e utilidade de dados, dependendo da escolha específica dos valores dos parâmetros, por exemplo, a taxa de permuta de dados.

Além destas recomendações, em qualquer caso, os detentores dos dados devem sempre estimar o risco de divulgação para seus conjuntos de dados originais, bem como os riscos de divulgação e utilidade dos dados para versões desidentificadas dos dados. Para alcançar bons resultados (ou seja, baixo risco de divulgação, alta utilidade de dados), é necessário desidentificar de forma iterativa, aplicando métodos diferentes, usando diferentes configurações de parâmetros, até que se consiga um equilíbrio adequado entre risco de divulgação e utilidade de dados.

1.5 – O fluxo de trabalho

O fluxo elaborado por Templ, Matthias (2017), na Figura 1 a seguir, mostra uma representação aproximada de um fluxo de trabalho comum para a aplicação do SDC



(Fig. 1 – Dos dados originais aos dados desidentificados divulgados: fluxo de trabalho simplificado. Retirado de "Templ, Matthias (2017). Statistical Disclosure Control for Microdata: Methods and Applications in R. Cap. 7 Practical Guidelines, p. 182. Springer. ISBN 978-3-319-50272-4", e publicado originalmente em "Templ, M. (2015). Quality indicators for statistical disclosure methods: A case study on the structural earnings survey. *Journal of Official Statistics*, 13(4), 737–761." sob os termos da Creative Commons Attribution License que permite qualquer uso, distribuição e reprodução em qualquer meio, desde que o(s) autor(es) e fonte originais sejam creditados).

As etapas de pré-processamento são cruciais, incluindo a discussão sobre possíveis Cenários de divulgação, a seleção de identificadores diretos, variáveis-chave e variáveis sensíveis, bem como para determinar um risco aceitável de divulgação e níveis de perda de utilidade de dados. Vários pontos cruciais são marcados na Figura 1 com números:

1 – O processo SDC real começa com a eliminação dos identificadores diretos;

2 – Para métodos não sintéticos, a seleção das variáveis-chave é de grande importância;

3 – Para variáveis-chave categóricas, antes de se aplicar qualquer técnica SDC, devem ser estimados os riscos de divulgação dos dados originais, incluindo os riscos de divulgação em nível individual e global. Isto permitirá a identificação de indivíduos com alto risco de divulgação, como aqueles que violam o *k*-anonimato (tipicamente 3-anonimato) ou observações com alto risco individual. Para variáveis-chave contínuas, por exemplo, a

estimativa de risco de divulgação com base na distância deve ser feita. Também o risco de divulgação global deve ser estimado usando uma abordagem de modelagem log-linear ou a soma dos riscos individuais;

4 – Dependendo do tipo de variáveis-chave (categóricas ou contínuas), as técnicas SDC podem então ser aplicadas. Para variáveis-chave categóricas, estas podem incluir uma supressão local - do inglês *local suppression* - para garantir o *k*-anonimato, ou quaisquer técnicas de troca. Para variáveis-chave contínuas, estas podem incluir o embaralhamento, microagregação ou adição de ruído, ou uma combinação dessas técnicas;

5 – Toda vez que uma técnica SDC é aplicada, as mesmas medidas de risco de divulgação devem ser repetidas e a extensão da perda de informações deve ser relatada; por exemplo, quantos valores foram suprimidos, quantas categorias foram combinadas ou como as estimativas sobre os indicadores mais importantes ficaram diferentes. Para variáveis-chave contínuas, o risco de divulgação é avaliado pela medida em que os conjuntos de dados originais e perturbados podem ser combinados. Desta forma, o risco de divulgação é, por padrão, de 100%, com relação ao conjunto de dados original.

Após a aplicação de qualquer método SDC, os riscos de divulgação podem ser medidos usando abordagens de vinculação de registros. A medida do risco de divulgação indica o percentual dos registros no conjunto de dados perturbado que pode ser vinculado àqueles nos dados originais. A medida de risco deve ser comparada e avaliada, conjuntamente com as medidas de perda de informação. Tanto para variáveis-chave categóricas como para as contínuas, a perda de informação deve ser quantificada não apenas por medidas diretas, mas também pela estimativa de indicadores ou modelos.

Os dados só estarão prontos para serem divulgados quando um nível aceitável de risco de divulgação tiver sido alcançado com uma perda mínima de informação. Caso contrário, técnicas alternativas de SDC devem ser aplicadas e/ou as mesmas técnicas devem ser repetidas com diferentes configurações de parâmetros.

A seguir, fornecemos algumas diretrizes práticas sobre questões comuns, tais como determinar variáveis-chave e avaliar níveis de riscos, e como selecionar os melhores métodos SDC para um determinado conjunto de dados.

1.5.1 – Como determinar as variáveis-chave

A maioria dos métodos de avaliação de risco de divulgação e SDC se baseia nas variáveis-chave selecionadas, que correspondem a certos cenários de divulgação. Na prática, determinar variáveis-chave é um desafio, já que não há regras definidas e qualquer variável potencialmente pertence a categoria de variáveis-chave, dependendo do cenário de divulgação.

A abordagem recomendada é considerar múltiplos cenários de divulgação e discutir com especialistas no assunto qual cenário é mais provável e realista. Um cenário comum é onde o intruso liga os dados divulgados com fontes de dados externas. Portanto, um passo importante é fazer um inventário de quais outras fontes de dados estão disponíveis e identificar variáveis que poderiam ser exploradas para se vincular aos dados divulgados. Além disso, variáveis sensíveis contendo informações confidenciais também devem ser identificadas de antemão. Em qualquer caso, é necessária uma estreita colaboração com o especialista da SDC e os especialistas no assunto para selecionar essas variáveis críticas.

1.5.2 – O Nível de Risco de Divulgação *versus* Perda de Informação

A avaliação da utilidade dos dados, especialmente para estimar indicadores ou modelos, requer conhecimento sobre os usuários finais dos dados desidentificados, como eles utilizarão os dados divulgados e, como resultado, quais informações devem ser preservadas. Se existe uma política de divulgação de microdados, o nível de risco aceitável varia para diferentes tipos de arquivos e condições de acesso. Por exemplo, os arquivos de uso público devem ter riscos de divulgação muito menores que os arquivos licenciados, cujo acesso é restrito a usuários específicos sujeitos a certos termos e condições. Além disso, um conjunto de dados contendo informações sensíveis, tais como informações sobre HIV/AIDS, pode exigir uma maior perturbação, em comparação com o que contém informações gerais e não sensíveis.

A melhor escolha dos estatísticos ou especialistas da SDC é consultar os especialistas em leis sobre privacidade e administração, discutir e fixar com eles o risco máximo tolerável de divulgação. Assim que esses números forem fixados, os especialistas da SDC podem agora modificar os dados para reduzir o risco abaixo do limite.

1.5.3 – Quais métodos SDC devem ser utilizados?

A força e a fraqueza de cada método SDC depende da estrutura do conjunto de dados e das variáveis-chave em consideração. A abordagem recomendada é a aplicação de diferentes métodos SDC com diferentes configurações de parâmetros de forma exploratória. A documentação do processo é, portanto, essencial para fazer comparações entre métodos e/ou parâmetros e para ajudar os produtores de dados a decidir sobre os níveis ótimos de perda de informações e risco de divulgação. Os parágrafos seguintes fornecem diretrizes gerais.

1.5.3.1 - Diretrizes Gerais para comparação de métodos

Para variáveis-chave categóricas, a recodificação é mais comumente usada. Se os riscos de divulgação permanecerem elevados após a recodificação, pode-se aplicar a supressão local para reduzir ainda mais o número de unidades de amostra. A recodificação deve ser aplicada de tal forma que uma mínima supressão local seja necessária posteriormente. Novamente, este processo de encontrar boas recodificações não se baseia em uma função de otimização matemática, mas sim em uma abordagem exploratória, iterativa, com conhecimento do assunto sobre o conjunto de dados a serem desidentificados.

A recodificação deve fazer sentido no que diz respeito ao conteúdo e aos sujeitos e, ao mesmo tempo, deve reduzir consideravelmente o risco. Por exemplo, a recodificação da idade em quatro faixas etárias ([0-19]; [20-39]; [40-59], [60-100]) não é recomendada quando a informação mais importante para a análise sociodemográfica é a idade. É melhor então não recodificar a idade, mas outras variáveis ou/e acrescentar algum ruído à idade. Outro exemplo é a modificação das informações regionais ou de localização. Quando o objetivo

dos pesquisadores é a análise espacial, eles não podem trabalhar com dados dos quais a informação regional é recodificada para regiões muito amplas. Por exemplo, para as estatísticas sobre os trabalhadores pendulares, são necessárias informações detalhadas sobre o local de vida e de trabalho. Aqui é melhor aplicar (com cuidado) procedimentos de troca como PRAM. Além disso, variáveis-chave categóricas podem ser incluídas onde não há necessidade de ter informações detalhadas; como por exemplo, quando o objetivo do pesquisador é comparar países.

Em geral, se um conjunto de dados tiver um grande número de variáveis-chave categóricas e/ou um grande número de categorias para determinadas variáveis-chave (por exemplo, variáveis de localização), a recodificação e supressão pode levar a uma perda excessiva de informações. Nessas situações, a PRAM pode ser uma abordagem mais vantajosa. A PRAM pode ser aplicada com ou sem recodificação prévia. Se a PRAM for aplicada após a recodificação, a matriz de transição deve especificar uma menor probabilidade de troca. Além disso, para variáveis sensíveis com alta *l*-diversidade, a recodificação e/ou PRAM são métodos úteis para aumentar o número de valores distintos de variáveis sensíveis, para cada grupo de observações compartilhando o mesmo padrão de variáveis-chave. No caso de muitas variáveis-chave ou no caso de fornecimento de arquivos de uso público com risco muito alto de divulgação, a geração de conjuntos de dados sintéticos é uma boa alternativa. Para variáveis-chave contínuas, a microagregação seria um método recomendado.

Para medir o risco de divulgação, em qualquer caso o *k*-anonimato deve ser verificado. Além disso, os riscos individuais e gerais devem ser estimados e outras supressões e recodificações devem ser feitas para as situações com alto risco na divulgação.

As medidas de risco global fornecem uma noção sobre o risco global de divulgação de um conjunto de dados e também são úteis na comparação entre conjuntos de dados desidentificados, ou seja, dados resultantes da aplicação de um conjunto de métodos SDC aos dados brutos com outro conjunto de métodos SDC aplicados ao mesmo conjunto de dados brutos. Ao produzir dados sintéticos, pode não ser necessário observar o risco de divulgação, uma vez que, por definição, ele deve ser muito baixo. No entanto, algumas correspondências com dados externos ou mesmo com os dados brutos da pesquisa podem ser feitas para avaliar o risco de divulgação.

Seção 2 – Instalação do *software* R e do pacote 'sdcMicro' e outros pacotes

Para os exemplos neste guia, recomendamos o uso do pacote *open source* (aberto) e gratuito para a SDC, chamado 'sdcMicro', dentro do *software* estatístico R.

O 'sdcMicro' é um pacote adicional ao *software* estatístico R. O pacote foi desenvolvido e é mantido por Matthias Templ, Alexander Kowarik e Bernhard Meindl. O *software* estatístico R e o pacote 'sdcMicro', assim como quaisquer outros pacotes necessários para o processo SDC, estão disponíveis gratuitamente nos espelhos da *Comprehensive R Archive Network* (CRAN)[2] .

O *software* está disponível para os sistemas operacionais Linux, Windows e Macintosh. Além da versão padrão do R, há uma interface mais amigável para o usuário do R, que é o RStudio.[3] .

É fortemente recomendado verificar regularmente por atualizações. Isto requer a instalação de uma nova versão para uma atualização do R; com o comando update.packages() ou usando as opções de menu em R ou RStudio pode-se atualizar os pacotes instalados.

O usuário precisa de algum conhecimento de R para usar o 'sdcMicro'. Está além do escopo deste guia ensinar o uso do R, porém fornecemos ao longo do guia exemplos de códigos sobre como implementar as rotinas necessárias no R. Além disso, através de um estudo de caso, demonstramos uma série de abordagens para o processo de desidentificação no R.

2.1. Passo a passo com o uso do R

Previamente é preciso que ao iniciar o R ou RStudio, se especifique quais embalagens estão

---

2 Disponível para download em <http://cran.r-project.org/>.

3 O RStudio também está disponível gratuitamente para Linux, Mac e Windows em: <http://www.rstudio.com>.

sendo utilizadas, carregando-as. Este carregamento de pacotes pode ser feito tanto com a função library() quanto com a função require().

```
> library(sdcMicro) # carregando o pacote sdcMicro

> require(sdcMicro) # carregando o pacote sdcMicro
```

Todos os pacotes e funções estão documentados. A maneira mais fácil de acessar a documentação de uma função específica é usar a 'ajuda integrada', que geralmente dá uma visão geral dos parâmetros das funções, assim como alguns exemplos.

A ajuda de uma função específica pode ser chamada por um ponto de interrogação seguido pelo nome da função, sem nenhum argumento.

```
> ?microaggregation # ajuda para a função de microagregação do pacote sdcMicro
```

A página de download de cada pacote no site da CRAN também fornece um manual de referência com uma visão geral completa das funções do pacote.

2.2 – Funções de leitura em R

O primeiro passo no processo SDC ao utilizar a sdcMicro é ler os dados em R e criar um quadro de dados – do inglês *data.frame*.

R é compatível com a maioria dos formatos de dados estatísticos e fornece funções de leitura para a maioria dos tipos de dados. Para essas funções de leitura, às vezes é necessário instalar pacotes adicionais e suas dependências em R.

O tamanho máximo de dados em R é tecnicamente restrito e dependente do sistema operacional. Alguns métodos SDC requerem longos tempos de computação para grandes conjuntos de dados.

2.3 – Valores Ausentes

A forma padrão de representação dos valores ausentes em R é pelo símbolo 'NA', que é diferente dos valores impossíveis, como a divisão por zero ou o log de um número negativo, que são representados pelo símbolo 'NaN'.

O valor 'NA' é usado tanto para variáveis numéricas como categóricas. Os valores suprimidos pela rotina localSuppression() também são substituídos pelo símbolo 'NA'.

Alguns conjuntos de dados e software estatístico podem usar valores diferentes para valores ausentes, como '999' ou strings. É possível incluir argumentos nas funções de leitura para especificar como os valores ausentes no conjunto de dados devem ser tratados e recodificar automaticamente os valores ausentes para 'NA'. Por exemplo, a função read.table() tem o argumento 'na.strings', que substitui as strings especificadas por valores 'NA'.

Os valores ausentes também podem ser recodificados após a leitura dos dados em R. Isto pode ser necessário se houver diversos códigos de valores em falta nos dados, diferentes códigos de valores em falta para diferentes variáveis ou a função de leitura para o tipo de dado não permitir especificar os códigos de valores em falta.

Ao preparar os dados no R, é importante recodificar quaisquer valores ausentes que não estejam codificados como 'NA' para 'NA', antes de iniciar o processo de desidentificação, de forma a assegurar a medição correta do risco (por exemplo, k-anonimato), bem como para assegurar que muitos dos métodos sejam corretamente aplicados aos dados.

2.4 – Classes no R

Todos os objetos no R são de uma classe específica, tais como *integer*, *character*, *matrix*, *factor* ou *data.frame*. A classe de um objeto é um atributo do qual o objeto herda. Para descobrir a classe de um objeto, pode-se usar a função class().

Funções em R podem requerer objetos ou argumentos de certas classes ou funções podem ter funcionalidades diferentes, dependendo da classe do argumento. Exemplos são as funções de escrita que requerem data.frames e a maioria das funções do pacote sdcMicro que requerem tanto data.frames quanto objetos da classe sdcMicroObj.

Uma classe importante definida e utilizada no pacote sdcMicro é a classe chamada sdcMicroObj. A funcionalidade das funções no pacote sdcMicro é diferente para os data.frames e objetos da classe sdcMicroObj. É fácil alterar o atributo de classe de um objeto com funções que começam com "as.", seguido pelo nome da classe - por exemplo, as.factor(), as.matrix(), as.data.frame().

2.5 – Objetos da classe sdcMicroObj

O pacote sdcMicro é construído em torno de objetos da classe sdcMicroObj. Cada membro desta classe tem uma certa estrutura com *slots* que contêm informações sobre o processo de desidentificação (ver Tabela 1 para uma descrição de todos os *slots*).

Antes de avaliar o risco e a utilidade e aplicar os métodos SDC, é recomendável criar um objeto da classe sdcMicro. A função usada para criar um objeto sdcMicro é createSdcObj().

A maioria das funções no pacote sdcMicro, tais como microaggregation() ou localSuppression(), usam automaticamente as informações necessárias, como, por exemplo, quase-identificadores, pesos de amostra do objeto sdcMicro, se aplicadas a um objeto da classe sdcMicro.

Os argumentos da função createSdcObj() permitem especificar o arquivo de dados original e categorizar as variáveis neste arquivo de dados antes do início do processo de desidentificação. Para isso, os cenários de divulgação já devem ter sido avaliados e os quase-identificadores selecionados. Além disso, deve-se garantir que não haja problemas com os dados, tais como variáveis contendo apenas valores ausentes.

Tabela 1 - Nomes de *slots* e descrição desses *slots* do objeto sdcMicroObj[4]

Slot	Conteúdo
origData	dados originais como especificado no argumento de dados da função createSdcObj()
keyVars	índices de colunas nos dados de origem com variáveis-chave categóricas especificadas
pramVars	índices de colunas nos dados de origem com variáveis PRAM especificadas
numVars	índices de colunas nos dados de origem com variáveis-chave numéricas especificadas
ghostVars	índices de colunas nos dados de origem com variáveis fantasmas especificadas
weightVar	índices de colunas nos dados de origem com variáveis de peso especificadas
hhId	índices de colunas nos dados de origem com variáveis de cluster/aglomeração especificadas
strataVar	índices de colunas nos dados de origem com variáveis de *strata*/divisões especificadas
sensibleVar	índices de colunas nos dados de origem com variáveis sensíveis para *l*-diversidade especificadas
manipKeyVars	variáveis-chave categóricas após aplicação de métodos SDC (vd. slot keyVars)
manipPramVars	variáveis PRAM após aplicação de métodos SDC (vd. slot pramVars)
manipNumVars	variáveis variáveis-chave numéricas após aplicação de métodos SDC (vd. slot numVars)
manipGhostVars	Variáveis fantasmas após aplicação de métodos SDC (vd. slot ghostVars)
manipStrataVar	Variáveis *strata*/divisões após aplicação de métodos SDC (vd. slot strataVar)
originalRisk	medidas originais de risco global e individual antes da desidentificação
risk	medidas de risco global e individual após a aplicação dos métodos SDC
utility	medidas de utilidade pública (il1 e eigen)
pram	detalhes da PRAM após a aplicação da PRAM
localSuppression	número de supressões locais por variável após supressão local
options	opções especificadas
additionalResults	resultados adicionais
set	lista de *slots* atualmente em uso (para uso interno)
prev	informação para desfazer um passo com a função undolast()
deletedVars	variáveis eliminadas (identificadores diretos)

Os nomes dos *slots* (Tabela 1) podem ser listados usando a função slotNames().

Note que nem todos os *slots* são utilizados em todos os casos. Alguns *slots* são preenchidos somente após a aplicação de determinados métodos, por exemplo, após avaliar uma medida de risco específica.

Já, alguns *slots* dos objetos podem ser acessados por funções de acesso (por exemplo, extractManipData() para extrair os dados desidentificados) ou funções de impressão (por exemplo, print()) com os argumentos apropriados.

[4] Tradução livre de tabela elaborada por "Templ, Matthias (2017). Statistical Disclosure Control for Microdata: Methods and Applications in R. Cap. 1 Software (1.4 Working with sdcMicro), p. 21. Springer. ISBN 978-3-319-50272-4").

O conteúdo de um *slot* também pode ser acessado diretamente com o operador '@' e o nome do *slot*. Esta funcionalidade pode ser prática para salvar resultados intermediários e comparar os resultados de diferentes métodos. Além disso, para mudanças manuais nos dados durante o processo SDC, tais como mudança de códigos de valores ausentes ou recodificação manual.

Dentro de cada *slot* há geralmente vários elementos. Seus nomes podem ser mostrados com a função names() e podem ser acessados com o operador '$'.

Há duas opções para salvar os resultados após a aplicação dos métodos SDC:

(1) Sobrescrever o objeto sdcMicro existente; ou,

(2) Criar um novo objeto sdcMicro.

O objeto original não será alterado e pode ser usado para comparar resultados. Isto é especialmente útil para comparar vários métodos e selecionar a melhor opção. Em ambos os casos, o resultado de qualquer função tem que ser reatribuído a um objeto com o operador '<-' ou '='. Se os resultados forem reatribuídos ao mesmo objeto sdcMicro, é possível desfazer a última etapa do processo SDC. Isto é útil quando se muda os parâmetros.

Os resultados da última etapa, entretanto, são perdidos depois de desfazer essa etapa. A função undolast() pode ser usada para dar apenas um passo atrás, não vários. O resultado também deve ser reatribuído ao mesmo objeto.

2.6 – Estrutura hierárquica

Se os dados tiverem uma estrutura hierárquica e algumas variáveis forem medidas no nível hierárquico superior e outras no nível inferior, o processo SDC deve ser adaptado de acordo. Um exemplo comum nos dados de pesquisa social são os conjuntos de dados com uma estrutura familiar ou domiciliar. As variáveis que são medidas no nível familiar são, por exemplo, renda familiar, tipo de casa e região. As variáveis medidas no nível individual são, por exemplo, idade, nível de educação e estado civil. Algumas variáveis são medidas no nível individual, mas são as mesmas para todos os membros da família em quase todos os

domicílios. Estas variáveis devem ser tratadas como medidas no nível domiciliar, da perspectiva da SDC.

O processo SDC deve ser dividido em duas etapas nos casos em que os dados tenham uma estrutura familiar ou domiciliar. Primeiro, as variáveis de nível superior (domiciliar) devem ser desidentificadas; posteriormente, as variáveis de nível superior, tratadas, devem ser fundidas com as variáveis individuais e desidentificadas em conjunto.

2.7 – Tempo de computação

Alguns métodos da SDC podem levar muito tempo para serem avaliados em termos de computação. Por exemplo, a supressão local com a função localSuppression() do pacote sdcMicro em R pode levar dias para ser executada em grandes conjuntos de dados de mais de 30.000 indivíduos que possuam muitos quase-identificadores categóricos.

Em geral, o tempo de computação é uma função dos seguintes fatores: o método SDC aplicado; tamanho dos dados, ou seja, número de observações, número de variáveis e o número de categorias ou níveis de fatores de cada variável categórica; complexidade dos dados (por exemplo, o número de diferentes combinações de valores de variáveis-chave nos dados); assim como as especificações do computador/servidor.

Dado o longo tempo de computação de alguns métodos, recomenda-se, sempre que possível, testar primeiro os métodos SDC em um subconjunto ou amostra dos microdados, e depois escolher os métodos SDC apropriados. O R fornece funções para selecionar subconjuntos a partir de um conjunto de dados. Após configurar o código, ele pode então ser executado em todo o conjunto de dados em um computador ou servidor poderoso.

Seção 3 – Aplicação de métodos SDC na desidentificação de microdados sintéticos

O estudo de caso aqui apresentado utiliza dados sintéticos que imitam as estruturas de pesquisas destinadas a medir a renda e o consumo de famílias, a participação da força de

trabalho e as características demográficas gerais.[5]

Este estudo de caso mostra um exemplo de como o processo de desidentificação pode ser conduzido utilizando a biblioteca de código aberto e gratuita sdcMicro no R, particularmente para um conjunto de dados com muitas variáveis contínuas.

Um roteiro R pronto para ser executado para este estudo de caso e o conjunto de dados também estão disponíveis para reproduzir os resultados e permitir ao usuário adaptar o código, em anexo. Pedaços deste código são apresentados nesta seção para ilustrar várias etapas do processo de desidentificação.

O estudo de caso apresentado segue 11 etapas do processo SDC, descritas a seguir.

3.1 – Etapa 1: Necessidade de controle de divulgação

As unidades estatísticas neste conjunto de dados são indivíduos e domicílios. A estrutura familiar fornece uma estrutura hierárquica nos dados, que deve ser levada em consideração ao medir o risco e selecionar métodos de desidentificação. Os dados contêm variáveis com informações demográficas, renda, despesas, variáveis educacionais e variáveis relacionadas ao *status* do trabalho do indivíduo. Essas variáveis incluem variáveis sensíveis e confidenciais.

O conjunto de dados é um exemplo de pesquisa social e, devido à natureza das unidades estatísticas e das variáveis, o controle da divulgação é necessário antes da divulgação dos microdados.

3.2 – Etapa 2: Preparação dos dados e exploração das características dos dados

O primeiro passo é explorar os dados. Para analisar os dados, primeiro temos que ler os

[5] Esses dados são públicos, disponíveis em <http://ihsn.org/sites/default/files/resources/case_studies_code_and_data.zip> pela International Household Survey Network (IHSN).

dados no R. Nesse caso, os dados estão salvos em um arquivo STATA (.dta file).[6]

Para ler arquivos STATA, precisamos carregar o pacote ‘foreign' do R. Também carregamos o pacote ‘sdcMicro’ e vários outros pacotes usados posteriormente para o cálculo das medidas de utilidade. Se estes pacotes ainda não estiverem instalados, você deve fazê-lo antes de tentar carregá-los. O código R para este estudo de caso demonstra como fazer isso.

```
# Load required packages

> library(foreign)   # for read/write function for STATA files

> library(sdcMicro)  # sdcMicro package with functions for the SDC process

> library(laeken)    # for GINI

> library(reldist)   # for GINI

> library(bootstrap) # for bootstrapping
```

Após definirmos o diretório de trabalho para o diretório onde o arquivo STATA está armazenado, carregamos os dados no objeto chamado file. Todas as saídas, a menos que especificado de outra forma, são salvas no diretório de trabalho.

```
#  Set working directory

> setwd("C:/WorldBank/CaseStudy/")

# Specify file name

> fname <- "case_1_data.dta"
```

6 O Stata é um software estatístico pago.

Read-in file

> file <- read.dta(fname, convert.factors = F) # factors as numeric code

A seguir, verificamos o número de variáveis, número de observações e nomes de variáveis

> dim(file) # Dimensions of file (observations, variables)

[1] 10574 68

> colnames(file) # Variable names

[1] "REGION" "DIST" "URBRUR" "WGTHH"

[5] "WGTPOP" "IDH" "IDP" "HHSIZE"

[9] "GENDER" "REL" "MARITAL" "AGEYRS"

[13] "AGEMTH" "RELIG" "ETHNICITY" "LANGUAGE"

[17] "MORBID" "MEASLES" "MEDATT" "CHWEIGHTKG"

89## [21] "CHHEIGHTCM" "ATSCHOOL" "EDUCY" "EDYRS"

[25] "EDYRSCURRAT" "SCHTYP" "LITERACY" "EMPTYP1"

[29] "UNEMP1" "INDUSTRY1" "EMPCAT1" "WHOURSWEEK1"

[33] "OWNHOUSE" "ROOF" "TOILET" "ELECTCON"

[37] "FUELCOOK" "WATER" "OWNAGLAND" "LANDSIZEHA"

[41] "OWNMOTORCYCLE" "CAR" "TV" "LIVESTOCK"

[45] "INCRMT" "INCWAGE" "INCBONSOCALL" "INCFARMBSN"

```
## [49] "INCNFARMBSN"   "INCRENT"       "INCFIN"        "INCPENSN"

## [53] "INCOTHER"      "INCTOTGROSSHH" "FARMEMP"       "THOUSEXP"

## [57] "TFOODEXP"      "TALCHEXP"      "TCLTHEXP"      "TFURNEXP"

## [61] "THLTHEXP"      "TTRANSEXP"     "TCOMMEXP"      "TRECEXP"

## [65] "TEDUEXP"       "TRESTHOTEXP"   "TMISCEXP"      "TANHHEXP"
```

O conjunto de dados tem 10.574 indivíduos em 2.000 domicílios e contém 68 variáveis. A pesquisa corresponde a uma população de cerca de 4,3 milhões de indivíduos, o que significa que a amostra é relativamente pequena e os pesos da amostra são altos. Isto tem um impacto sobre o risco de divulgação, como veremos nos Passos 6a e 6b.

Para obter uma visão geral dos valores das variáveis, usamos tabulações e tabulações cruzadas para variáveis categóricas e estatísticas resumidas para variáveis contínuas. Para incluir o número de valores ausentes (NA ou outros), usamos a opção useNA = "ifany", na função table().

Na Tabela 2, as variáveis do conjunto de dados são listadas conjuntamente com descrições concisas das variáveis, o nível no qual elas são coletadas (individual (IND), domiciliar (DM)), o tipo de medida (contínua, semicontínua, categórica) e as faixas de valores.

As variáveis foram pré-selecionadas (68 de 112) com base em sua relevância para os usuários dos dados. Isto permite reduzir o número total de variáveis a serem consideradas no processo de desidentificação, facilitando o processo.

Os valores numéricos para muitas das variáveis categóricas são códigos que se referem a valores, por exemplo, na variável URBRUR, 1 representa o rural e 2 o urbano. Mais informações sobre os significados dos valores codificados das variáveis categóricas estão disponíveis no código R para este estudo de caso.

Tabela 2 - Descrição das variáveis do subconjunto sintético de dados do estudo de caso.

No.	Nome da Variável	Descrição	Nível	Tipo	Valores
1	IDH	Identificação do lar	DM	.	1-2,000
2	IDP	Identificação individual	IND	.	1-33
3	REGION	Região	DM	categórica	1-6
4	DISTRICT	Distrito	DM	categórica	101-1105
5	URBRUR	Área de residência	DM	categórica	1, 2
6	WGTHH	Coeficiente de ponderação individual	DM	peso	31,2-8495
7	WGTPOP	Coeficiente de ponderação da população	IND	peso	45,8-93452,2
8	HHSIZE	Tamanho do lar	DM	semi-contente	1-33
9	GENDER	Gênero	IND	categórica	0, 1
10	REL	Relação com o chefe de família	IND	categórica	1-9
11	MARITAL	Estado civil	IND	categórica	1-6
12	AGEYRS	Idade em anos completos	IND	semi-contínuo	0-95 (menos de 1, incrementos de 1/12 anos)
13	AGEMTH	Idade da criança em anos completos	IND	semi-contínuo	1-1140
14	RELIG	Religião do chefe de família	DM	categórica	1, 5-7, 9
15	ETHNICITY	Etnia do chefe de família	DM	categórica	todos os valores em falta
16	LANGUAGE	Idioma do chefe de família	DM	categórica	todos os valores em falta
17	MORBID	A morbidez dura x semanas	IND	categórica	0, 1
18	MEASLES	Criança imunizada contra o sarampo	IND	categórica	0, 1, 9
19	MEDATT	Procura de atendimento médico	IND	categórica	0, 1
20	CHWEIGHTKG	Peso da criança (Kg)	IND	contínuo	2 – 26,5
21	CHHEIGHTCM	Altura da criança (cms)	IND	contínuo	7 – 140
22	ATSCHOOL	Atualmente matriculados na escola	IND	categórica	0, 1
23	EDUCY	O mais alto nível de educação frequentado	IND	categórica	1-6
24	EDYEARS	Anos de educação	IND	semi-contínuo	0-18
25	EDYRSCURRAT	Anos de educação para os matriculados atualmente	IND	semi-contínuo	1-18
26	SCHTYP	Tipo de escola que frequenta	IND	categórica	1-3, 9
27	LITERACY	Alfabetização	IND	categórica	1-3
28	EMPTYP1	Tipo de emprego	IND	categórica	1-9
29	UNEMP1	Desempregado	IND	categórica	0, 1
30	INDUSTRY1	Classificação do setor (1 dígito)	IND	categórica	1-10

31	EMPCAT1	Categorias de emprego	IND	categórica	11, 12, 13, 14, 21, 22
32	WHOURSLASTWEEK1	Horas trabalhadas na semana passada	IND	contínuo	0-154
33	OWNHOUSE	Posse de moradia	DM	categórica	0, 1
34	ROOF	Principal material utilizado para telhado	IND	categórica	1-5, 9
35	TOILET	Instalações sanitárias principais	DM	categórica	1-4, 9
36	ELECTCON	Eletricidade	DM	categórica	0-3
37	FUELCOOK	Combustível principal de cozinha	DM	categórica	1-5, 9
38	WATER	Principal fonte de água	DM	categórica	1-9
39	OWNAGLAND	Propriedade de terras agrícolas	DM	categórica	1-3
40	LANDSIZEHA	Tamanho da terra de propriedade da família (ha) (agrícola e não agrícola)	DM	contínuo	0-1214
41	OWNMOTORCYCLE	Posse de motocicleta	DM	categórica	0, 1
42	CAR	Propriedade do carro	DM	categórica	0, 1
43	TV	Propriedade da televisão	DM	categórica	0, 1
44	LIFESTOCK	Número de animais de grande porte de propriedade	DM	semi-contínuo	0-25
45	INCRMT	Renda - Remessas	DM	contínuo	0 – 3e+05
46	INCWAGE	Renda - Salários e vencimentos	DM	contínuo	0 – 683922
47	INCBONSOCAL	Renda - Bônus e abonos sociais derivados de empregos assalariados	DM	contínuo	0 – 60000
48	INCFARMBSN	Renda - Renda bruta das empresas agrícolas domésticas	DM	contínuo	0 – 165400
49	INCNFARMBSN	Renda - Renda bruta dos negócios domésticos não agrícolas	DM	contínuo	0 – 4e+05
50	INCRENT	Renda - Aluguel	DM	contínuo	0 – 120.000
51	INCFIN	Renda - Financeira	DM	contínuo	0 – 14.400
52	INCPENSN	Renda - Pensões/assistência social	DM	contínuo	0 – 60.000
53	INCOTHER	Renda - Outros	DM	contínuo	0 – 82.300
54	INCTOTGROSHH	Renda - Total	DM	contínuo	5000 – 683.922
55	FARMEMP	Emprego em fazenda	DM	categórica	0, 1
56	TFOODEXP	Gastos totais com alimentos	DM	contínuo	0 – 197.544
57	TALCHEXP	Gastos totais com bebidas alcoólicas, tabaco e narcóticos	DM	contínuo	0 – 127.920

58	TCLTHEXP	Despesas totais com vestuário	DM contínuo	0 – 85.280
59	THOUSEXP	Despesas totais com moradia	DM contínuo	0 – 28.400
60	TFURNEXP	Gastos totais com mobiliário	DM contínuo	0 – 17.778
61	THLTHEXP	Gastos totais com saúde	DM contínuo	0 – 49.653
62	TTRANSEXP	Despesas totais com transporte	DM contínuo	0 – 91.920
63	TCOMMEXP	Gastos totais com comunicação	DM contínuo	0 – 34.000
64	TRECEXP	Gastos totais com recreação	DM contínuo	0 – 15.876
65	TEDUEXP	Gastos totais com educação	DM contínuo	0 – 240.309
66	TRESHOTEXP	Despesas totais em restaurantes e hotéis	DM contínuo	0 – 63.700
67	TMISCEXP	Gastos totais com despesas diversas	DM contínuo	0 – 67.416
68	TANHHEXP	Gastos domésticos nominais totais anuais	DM contínuo	498 – 353.230

(Dados públicos, disponíveis em <http://ihsn.org/sites/default/files/resources/case_studies_code_and_data.zip> pela International Household Survey Network (IHSN))

Analisando a Tabela 2 acima, identificamos as seguintes variáveis sensíveis nos dados: ETHNICITY, RELIGION, assim como as variáveis relacionadas ao *status* da força de trabalho do indivíduo e as variáveis contendo informações sobre renda e despesas do domicílio.

As variáveis identificadas como sensíveis podem variar, dependendo do país e/ou do conjunto de dados.

O conjunto de dados do estudo de caso não tem identificadores diretos que, se estivessem presentes, precisariam ser removidos nesta fase. Exemplos de identificadores diretos seriam nomes, números de telefone, coordenadas de localização geográfica, etc.

```
# tabulation of variable GENDER (sex, categorical)

> table(file$GENDER, useNA = "ifany")

##    0    1
```

```
## 5448 5126
```

```
# summary statistics for variable TANHHEXP (total annual household expenditures,
# continuous)
> summary(file$TANHHEXP)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     498   15550   17290   28560   29720  353200
```

É sempre importante assegurar que as relações entre as variáveis nos dados sejam preservadas durante o processo de desidentificação. Assim, é necessário explorar e tomar nota dessas relações antes de iniciar a desidentificação.

Na etapa final do processo de desidentificação, deve ser realizada uma auditoria, utilizando estes resultados iniciais, para verificar se estas relações são mantidas no conjunto de dados anonimizados.

Em nosso conjunto de dados, identificamos várias relações entre as variáveis que precisam ser preservadas durante o processo de desidentificação. Por exemplo, as variáveis TANHHEXP e INCTOTGROSSHH representam a despesa nominal anual total dos domicílios e a renda bruta anual total dos domicílios, respectivamente, e estas variáveis são agregações dos componentes de renda e despesa existentes no conjunto de dados.

As variáveis relacionadas à educação estão disponíveis apenas para indivíduos nos grupos etários apropriados e ausentes para outros indivíduos.

Fazemos uma observação semelhante para variáveis relacionadas às crianças, tais como altura, peso e idade, em meses.

Além disso, as variáveis de nível doméstico têm os mesmos valores para todos os membros

de qualquer família em particular. O valor do tamanho do domicílio corresponde ao número real de indivíduos pertencentes a esse domicílio no conjunto de dados. Conforme prosseguimos, temos que cuidar para que essas relações e estruturas sejam preservadas no processo de desidentificação.

Ao tabular as variáveis, notamos que as variáveis RELIG, EMPTYP1 e LIVESTOCK têm códigos de valor faltantes diferentes do código de valor ausente padrão do R, NA.

Antes de prosseguir, precisamos recodificá-las para NA para que o R as interprete corretamente.

Os códigos de valores ausentes são, respectivamente, 99999, 99 e 9999 para estas três variáveis. Estes são códigos de valores ausentes genuínos e não causados pelo fato de as variáveis não serem aplicáveis ao indivíduo. A seguir mostramos como fazer estas mudanças:

```
# Set different NA codes to R missing value NA

> file[,'RELIG'][file[,'RELIG'] == 99999]        <- NA

> file[,'EMPTYP1'][file[,'EMPTYP1'] == 99]       <- NA

> file[,'LIVESTOCK'][file[,'LIVESTOCK'] == 9999] <- NA
```

Observamos também que as variáveis LANGUAGE e ETHNICITY têm apenas valores ausentes.

As variáveis que contêm apenas valores ausentes devem ser removidas do conjunto de dados nesta fase e excluídas do processo de desidentificação.

A remoção destas variáveis não significa perda de dados ou redução da utilidade dos dados, uma vez que estas variáveis não possuíam nenhuma informação. Entretanto, é necessário removê-las, pois mantê-las pode levar a erros em alguns dos métodos de desidentificação no

R. É sempre possível adicionar estas variáveis de volta ao conjunto de dados a serem liberadas ao final do processo de desidentificação.

É útil reduzir o conjunto de dados a essas variáveis e registros relevantes para o processo de desidentificação. Isto garante os melhores resultados no R e menos erros.

A seguir, removemos as variáveis que contêm todos os valores ausentes:

```
# Drop variables containing only missings

> file <- file[,!names(file) %in% c('LANGUAGE', 'ETHNICITY')]
```

No caso, vamos supor que os dados são coletados em uma pesquisa que utiliza uma amostragem simples dos domicílios.

Os dados contêm dois coeficientes de peso: WGTHH e WGTPOP. A relação entre os pesos é WGTPOP = WGTHH * HHSIZE.

WGTPOP é o peso da amostragem para os domicílios e WGTHH é o peso da amostragem para os indivíduos a serem usados para os cálculos de risco de divulgação.

WGTHH é usado para calcular indicadores de nível individual (como educação) e WGTPOP é usado para indicadores de nível populacional (como indicadores de renda). Não há variáveis de estratos disponíveis nos dados. Usaremos WGTPOP para a anonimização das variáveis domésticas e WGTHH para a anonimização das variáveis de nível individual.

3.3 – Etapa 3: Tipo de liberação

Neste estudo de caso, partimos do princípio de que os dados de pesquisa só estarão disponíveis, mediante requisição de pesquisadores para que sejam credenciados e suas propostas de pesquisa previamente aprovadas por um comitê de ética. Portanto, estes dados

integrarão um banco de dados de acesso restrito. Logo, o nível de risco aceito é maior e um conjunto mais amplo de variáveis pode ser liberado ao compararmos com a liberação de dados de um arquivo de uso público.

Como não temos uma visão geral das exigências de todos os usuários, restringimos as medidas de utilidade a um número selecionado de usos de dados (ver Passo 5).

3.4 – Etapa 4: Cenários de intrusos e escolha das variáveis-chave

Em seguida, analisamos possíveis cenários de intrusão e selecionamos quase-identificadores ou variáveis-chave com base nesses cenários.

Como o conjunto de dados utilizado neste estudo de caso é um conjunto de dados sintético que não provém de um país existente (e, portanto, não temos informações sobre fontes de dados externas disponíveis para possíveis intrusos), não é possível definir cenários exatos de divulgação.

Por esta razão, esboçamos cenários de intrusos para este conjunto de dados sintético com base em algumas suposições hipotéticas sobre a disponibilidade de fontes de dados externas. Consideramos dois tipos de cenários de divulgação: (1) correspondência com outros conjuntos de dados disponíveis publicamente; e, (2) reconhecimento espontâneo. Descreveremos os dois cenários nos dois parágrafos seguintes.

Para fins de ilustração, assumimos que os registros de população estão disponíveis com as variáveis demográficas: sexo, idade, local de residência (região, urbano/rural), religião e outras variáveis tais como estado civil e variáveis relacionadas à educação e *status* profissional, que também estão presentes em nosso conjunto de dados. Além disso, assumimos que existe um registro cadastral publicamente disponível sobre a propriedade da terra.

Com base nesta análise das fontes de dados disponíveis, selecionamos as variáveis REGION, URBRUR, HHSIZE, OWNAGLAND, RELIG, GENDER, REL (relação com o chefe de família), MARITAL (estado civil), AGEYRS, INDUSTRY1 e duas variáveis relativas à

frequência escolar como quase-identificadores categóricos, as variáveis de gastos e rendimentos, bem como LANDSIZEHA como quase-identificadores contínuos.

Estas variáveis podem permitir que um intruso reidentifique um indivíduo ou um agregado familiar no conjunto de dados, combinando com outros conjuntos de dados disponíveis.

A decisão de disponibilizar o conjunto de dados como dados de pesquisa significa que o nível de desidentificação será relativamente baixo e, consequentemente, as variáveis serão mais detalhadas e um cenário de reconhecimento espontâneo torna-se nossa principal preocupação.

Portanto, devemos verificar se há combinações raras ou padrões incomuns nas variáveis.

As variáveis que podem levar ao reconhecimento espontâneo em nossa amostra são, entre outras, HHSIZE (tamanho do agregado familiar), LANDSIZEHA, bem como variáveis de renda e despesa.

Grandes residências e grandes propriedades fundiárias são facilmente identificáveis. O mesmo se aplica a valores extremos de variáveis de riqueza e despesas, especialmente quando combinadas com outras variáveis de identificação, como região. Pode haver apenas uma ou poucas famílias em uma determinada região com alta renda, como, por exemplo, o médico do local.

Variáveis que são facilmente observáveis e conhecidas por vizinhos como ROOF, TOILET, WATER, ELECTCON, FUELCOOK, OWNMOTORCYCLE, CAR, TV e LIVESTOCK também podem precisar de proteção dependendo do que se destaca na comunidade, uma vez que um pesquisador pode ser capaz de identificar pessoas(s) que ele conhece. Isto é chamado de cenário de vizinhança intrometida.

3.5 – Etapa 5: Utilização da chave de dados e seleção de medidas de utilidade

Neste estudo de caso, nosso objetivo é criar um arquivo de dados que forneça informações suficientes para os pesquisadores credenciados. Sabemos que o principal uso desses dados será avaliar indicadores relacionados à renda e à desigualdade. Exemplos desses são o

coeficiente GINI e os indicadores sobre que parcela da renda é gasta em que tipo de despesas.

Além disso, nos concentramos em alguns indicadores de educação, são eles: (a) Curvas de Lorenz para gastos totais, (b) Média mensal dos gastos totais *per capita* por área de residência, (c) Participação média dos componentes nas despesas, (d) Renda total média mensal per capita por área de residência, (e) Participação média dos componentes na renda e (f) Matrícula líquida no ensino primário por gênero.

Além dessas medidas de utilidade, que são específicas aos usos dos dados, também fazemos verificações padrão, tais como comparação de tabulações, tabulações cruzadas e estatísticas resumidas antes e depois da desidentificação. Grandes diferenças reduziriam a credibilidade do conjunto de dados desidentificado.

Há, nos dados sintéticos, coleta de dados sobre indivíduos em residências. A estrutura domiciliar é importante para os usuários de dados e deve ser considerada na avaliação de risco. Uma vez que algumas variáveis são medidas no nível doméstico e, portanto, têm valores idênticos para os membros da família, essas devem ser tratadas da mesma maneira para cada membro da família.

Portanto, primeiro desidentificamos apenas as variáveis do agregado familiar. Depois disso, nós as fundimos com as variáveis de nível individual e depois anonimizamos as variáveis de nível individual e de nível doméstico em conjunto.

Como os dados têm uma estrutura hierárquica, os Passos 6 a 10 são repetidos duas vezes: Os Passos 6a a 10a são para as variáveis de nível doméstico e os Passos 6b a 10b para o conjunto de dados combinados.

Desta forma, garantimos que os valores das variáveis de nível doméstico permaneçam consistentes entre os membros do domicílio para cada domicílio e a estrutura do domicílio não possa ser usada para reidentificar os indivíduos.

Antes de continuarmos para o Passo 6a, selecionamos as variáveis-chave categóricas, variáveis-chave contínuas e quaisquer variáveis selecionadas para uso em rotinas PRAM, bem como os pesos de amostragem relativas a domicílio.

Extraímos estas variáveis domésticas selecionadas e os domicílios do conjunto de dados e as salvamos em um novo conjunto de dados de nome fileHH. A escolha das variáveis PRAM é explicada com mais detalhes no Passo 8a. A seguir, mostramos como estas etapas são feitas no R.

```
### Select variables (household level)

# Key variables (household level)

> selectedKeyVarsHH = c('URBRUR', 'REGION', 'HHSIZE', 'OWNHOUSE', 'OWNAGLAND',
'RELIG')


# Changing variables to class factor

> file$URBRUR    <- as.factor(file$URBRUR)

> file$REGION    <- as.factor(file$REGION)

> file$OWNHOUSE  <- as.factor(file$OWNHOUSE)

> file$OWNAGLAND <- as.factor(file$OWNAGLAND)

> file$RELIG     <- as.factor(file$RELIG)


# Numerical variables

> numVarsHH = c('LANDSIZEHA', 'TANHHEXP', 'TFOODEXP', 'TALCHEXP', 'TCLTHEXP',
        'THOUSEXP', 'TFURNEXP', 'THLTHEXP', 'TTRANSEXP',  'TCOMMEXP', 'TRECEXP',
        'TEDUEXP',  'TRESHOTEXP',  'TMISCEXP',  'INCTOTGROSSHH',  'INCRMT',
        'INCWAGE',  'INCFARMBSN',  'INCNFARMBSN',  'INCRENT',  'INCFIN',  'INCPENSN',
        'INCOTHER')
```

```
# PRAM variables

> pramVarsHH = c('ROOF', 'TOILET', 'WATER', 'ELECTCON', 'FUELCOOK',
    'OWNMOTORCYCLE', 'CAR', 'TV', 'LIVESTOCK')


# sample weight (WGTPOP) (household)

> weightVarHH = c('WGTPOP')


# All household level variables

> HHVars <- c('HID', selectedKeyVarsHH, pramVarsHH, numVarsHH, weightVarHH)
```

Em seguida, extraímos estas variáveis do arquivo, o *data.frame* em R que contém todas as variáveis. Cada domicílio tem o mesmo número de entradas que tem membros (por exemplo, uma família de três será repetida três vezes no fileHH).

Antes de analisar as variáveis de nível doméstico, selecionamos apenas uma entrada por domicílio, como a seguir.

```
# Create subset of file with households and HH variables

> fileHH <- file[,HHVars]


# Remove duplicated rows based on IDH, select uniques,

# one row per household in fileHH
```

```
> fileHH <- fileHH[which(!duplicated(fileHH$IDH)),]

> dim(fileHH)

## [1] 2000   39
```

O fileHH contém 2.000 domicílios e 39 variáveis.

Estamos agora prontos para criar nosso objeto sdcMicro com as variáveis correspondentes que selecionamos.

Para nosso estudo de caso, criaremos um objeto sdcMicro chamado sdcHH com base nos dados do fileHH, que utilizaremos para os passos 6a – 10a.

```
# Create initial SDC object for household level variables

> sdcHH <- createSdcObj(dat = fileHH, keyVars = selectedKeyVarsHH, pramVars =
      pramVarsHH, weightVar = weightVarHH, numVars = numVarsHH)

> numHH <- length(fileHH[,1]) # number of households
```

3.6 – Etapa 6a: Avaliar o risco de divulgação (nível domiciliar)

Como primeira medida, avaliamos o número de domicílios que violam o k-anonimato nos limiares 2, 3 e 5.

```
# Number of observations violating k-anonymity (thresholds 2 and 3)
```

```
> print(sdcHH)

## Infos on 2/3-Anonymity:

##

## Number of observations violating

##  - 2-anonymity: 103

##  - 3-anonymity: 229

##

## Percentage of observations violating

##  - 2-anonymity: 5.150 %

##  - 3-anonymity: 11.450 %

-----------------------------------------------------------------------


# Calculate sample frequencies and count number of obs. violating k(5) - anonymity

> kAnon5 <- sum(sdcHH@risk$individual[,2] < 5)

> kAnon5

## [1] 489


# As percentage of total

> kAnon5 / numHH

## [1] 0.2445
```

A função print() na sdcMicro mostra apenas os valores para os limiares 2 e 3. Valores para outros limiares podem ser calculados manualmente somando as frequências menores que o limiar de k-anonimato, como mostrado acima para o limiar 5.

Muitas vezes é útil visualizar os valores para o(s) domicílio(s) que violam o k-anonimato, o que pode ajudar a esclarecer quais variáveis causam a singularidade destes domicílios. Isto pode então ser usado mais tarde ao escolher os métodos SDC apropriados.

A seguir mostramos como avaliar os valores dos domicílios que violam o 3- e 5-anonimato.

```
# Show values of key variable of records that violate k-anonymity

> fileHH[sdcHH@risk$individual[,2] < 3, selectedKeyVarsHH] # for 3-anonymity

> fileHH[sdcHH@risk$individual[,2] < 5, selectedKeyVarsHH] # for 5-anonymity
```

Parece que entre as variáveis-chave categóricas, a variável HHSIZE é responsável por muitas das combinações únicas e a origem de grande parte do risco. Tendo determinado isto, podemos assinalar HHSIZE como uma possível primeira variável a ser tratada para obter o nível de risco requerido.

Na prática, com uma variável como HHSIZE, isto provavelmente envolverá a remoção de grandes residências do conjunto de dados a ser liberado. Como já explicado, a recodificação e a supressão local não são opções válidas para a variável HHSIZE.

As frequências de tamanho doméstico mostram que há poucos domicílios com mais de 13 membros domiciliares. Isso torna esses domicílios facilmente identificáveis com base no número de membros do domicílio e com alto risco de reidentificação, também no contexto do cenário do vizinho intrometido.

Também avaliamos o risco de divulgação das variáveis categóricas com as medidas de risco individual e global. No fileHH cada entrada representa um domicílio. Portanto, usamos aqui o risco individual não hierárquico, onde o indivíduo se refere neste caso a um domicílio.

O fileHH contém apenas domicílios e não tem estrutura hierárquica.

No passo 6b, avaliamos o risco hierárquico no arquivo, o conjunto de dados que contém tanto os domicílios quanto os indivíduos.

As medidas de risco individual e global levam automaticamente em consideração os pesos de amostragem dos domicílios. Em nosso arquivo, a medida de risco global calculada usando as variáveis-chave escolhidas é de 0,05%. Esta porcentagem é extremamente baixa e corresponde a 1,03 reidentificações esperadas. A seguir mostramos como imprimir a medida de risco global.

```
> print(sdcHH, "risk")

## Risk measures:

##

## Number of observations with higher risk than the main part of the data: 0

## Expected number of re-identifications: 1.03 (0.05 %)
```

Este valor baixo pode ser explicado pelo tamanho relativamente pequeno da amostra de 0,25% da população total. Além disso, deve-se ter em mente que esta medida de risco se baseia apenas nos quase-identificadores categóricos no nível doméstico.

A medida de risco global não fornece informações sobre a propagação das medidas de risco individuais. Pode haver domicílios com risco relativamente alto, enquanto o risco global (médio) é baixo.

Portanto, é útil, como próximo passo, inspecionar as observações com risco relativamente alto. O risco mais alto é de 5,5% e apenas 14 residências têm risco maior que 1%.

A seguir mostramos como exibir os domicílios com risco acima de um determinado limite. Aqui o limiar é de 0,01 (1%).

```
# Observations with risk above certain threshold (0.01)

> fileHH[sdcHH@risk$individual[, "risk"] > 0.01,]
```

Como as variáveis-chave selecionadas no âmbito doméstico são tanto categóricas quanto numéricas, as medidas de risco individuais e globais baseadas em contagens de frequência não refletem completamente o risco de divulgação de todo o conjunto de dados. Tanto as variáveis-chave categóricas quanto as contínuas são importantes para os usuários de dados, portanto, opções como recodificar as variáveis contínuas (por exemplo, criando fatias de renda e despesa) para torná-las todas categóricas, provavelmente não satisfarão as necessidades do usuário de dados. Por isso, evitamos a recodificação de variáveis contínuas e avaliamos o risco de divulgação das variáveis categóricas e contínuas separadamente.

Esta abordagem pode ser parcialmente justificada pelo fato de que qualquer potencial correspondência com fontes de dados externas para as variáveis contínuas e categóricas está disponível a partir de diferentes fontes de dados externas e, como tal, não será usada simultaneamente para correspondência.

Para medir o risco das variáveis contínuas, usamos uma medida de intervalo, que mede o número de valores anonimizados que estão muito próximos de seus valores originais.

Essa medida é uma medida *ex-post*, o que significa que o risco só pode ser avaliado após a

desidentificação e mede se o transtorno é suficientemente grande. Como é uma medida *ex-post*, só podemos avaliá-la na etapa 9ª, após as variáveis terem sido tratadas.

Avaliar esta medida com base nos dados originais levaria a um risco de 100%, já que todos os valores estariam muito próximos dos valores originais, não importando quão pequenos fossem os intervalos escolhidos.

Também analisamos a distribuição da LANDSIZEHA.

Na variável LANDSIZEHA, os valores altos são raros e podem levar a uma reidentificação. Um exemplo é um grande proprietário de terras em uma região específica. Para avaliar a distribuição da variável LANDSIZEHA, olhamos para os percentis. Cada percentil representa aproximadamente 20 residências. Além disso, observamos os valores dos 50 maiores lotes. A seguir mostramos como usar o R para exibir os quintis e as maiores parcelas de terreno:

```
# 1st - 100th percentiles of land size

> quantile(fileHH$LANDSIZEHA, probs = (1:100)/100, na.rm= TRUE)


# Values of landsize for largest 50 plots

> tail(sort(fileHH$LANDSIZEHA), n = 50)
```

Com base nestes valores, concluímos que valores de LANDSIZEHA acima de 40 tornam o agregado familiar muito identificável. Estes grandes lares e as famílias com grandes parcelas de terra precisam de proteção extra, como discutido no Passo 8a.

3.7 – Etapa 7a: Avaliação das medidas de utilidade (nível doméstico)

A utilidade dos dados não depende apenas das variáveis de nível doméstico, mas da combinação de variáveis de nível doméstico e de nível individual. Logo, não é útil avaliar todas as medidas de utilidade selecionadas na Etapa 5, ou seja, antes de anonimizar as variáveis de âmbito individual. Restringimos a medida inicial de utilidade àquelas medidas que se baseiam exclusivamente nas variáveis de âmbito doméstico.

Em nosso conjunto de dados, estas são as medidas relacionadas às receitas e despesas e suas distribuições. Os resultados são apresentados na etapa 10a, conjuntamente com os resultados, após a desidentificação, permitindo, assim, uma comparação direta. Se, após a próxima etapa de desidentificação, parecer que a utilidade dos dados foi significativamente reduzida pela supressão de algumas variáveis de doméstico, podemos voltar a esta etapa.

3.8 – Etapa 8a: Escolha e aplicação dos métodos SDC (variáveis domiciliares)

Esta etapa é dividida na desidentificação da variável HHSIZE, pois este é um caso especial: A desidentificação dos outros quase-identificadores categóricos selecionados e a desidentificação dos quase-identificadores contínuos selecionados.

A variável HHSIZE representa um problema para a desidentificação do arquivo, uma vez que uma simples contagem do número de repetições da identidade do domicílio, que é o número de habitantes do domicílio, permitiria a reconstrução desta variável. O número de domicílios para cada tamanho maior que 13 é de 6 ou menos e pode ser considerado como sendo um valor extremo com maior risco de reidentificação, como discutido na etapa 6a.

Uma maneira de lidar com isso é remover todos os domicílios de tamanho 14 ou maior do conjunto de dados. A remoção de 29 residências de tamanho 14 ou maior, reduz o número de violações de 2-anonimato em 18, de 3-anonimato em 26 e de 5-anonimato em 29.

Isto significa que todos os domicílios removidos violaram o 5-anonimato devido ao valor da variável HHSIZE e muitos deles o 2 ou 3-anonimato.

Além disso, o risco individual médio entre os 29 domicílios é de 0,15%, o que é quase três

vezes maior do que o risco individual médio de todos os domicílios. O impacto na medida de risco global da remoção desses 29 domicílios é, no entanto, limitado, devido ao número relativamente pequeno de domicílios removidos em comparação com o número total de 2.000 domicílios.

A remoção dos domicílios se faz principalmente para proteger esses domicílios específicos, não para reduzir o risco global. Mudanças, como a remoção de registros, não podem ser feitas diretamente no objeto sdcMicro.

A seguir ilustramos uma maneira de remover residências e recriar o objeto da sdcMicro:

```
# Tabulation of variable HHSIZE

> table(sdcHH@manipKeyVars$HHSIZE)


# Remove large households (14 or more household members) from file and fileHH

> file <- file[!file,'HHSIZE'] >= 14,]

> fileHHnew <- fileHH[!fileHH,'HHSIZE'] >= 14,]


# Create new sdcMicro object based on the file without the removed households

> sdcHH <- createSdcObj(dat=fileHHnew, keyVars=selectedKeyVarsHH,
        pramVars=pramVarsHH, weightVar=weightVarHH, numVars = numVarsHH)
```

Estamos agora prontos para passar à escolha dos métodos SDC para as variáveis categóricas no nível doméstico em nosso conjunto de dados.

Como observado em nossa discussão dos métodos, a aplicação de métodos perturbadores e de supressão local pode levar a uma grande perda de utilidade. A abordagem comum é aplicar a recodificação na maior extensão possível como primeira abordagem, para alcançar um nível de risco prescrito e reduzir o número de supressões necessárias. Somente depois disso, métodos como a supressão local devem ser aplicados.

Se esta abordagem ainda não atingir o resultado desejado, podemos considerar métodos perturbadores.

Como o arquivo deve ser liberado como dados de pesquisa em um banco de dados de acesso restrito, podemos manter um nível mais alto de detalhes nos dados.

As variáveis-chave categóricas selecionadas no nível doméstico não são adequadas para recodificação neste ponto. Devido ao risco relativamente baixo de reidentificação, com base nas cinco variáveis categóricas selecionadas no âmbito doméstico, é possível, neste caso, usar uma opção como a supressão local para atingir nosso nível de risco desejado.

A aplicação da supressão local, quando o risco inicial é relativamente baixo, provavelmente levará apenas à supressão de poucas observações e assim limitará a perda de utilidade. Se, entretanto, os dados tivessem sido medidos para ter um risco relativamente alto, então a aplicação da supressão local, sem recodificação prévia, provavelmente resultaria em um grande número de supressões e maior perda de informações.

Esforços de tal recodificação devem ser feitos primeiro antes da supressão, nos casos em que o risco é inicialmente medido como alto. A recodificação reduzirá o risco com pouca perda de informação e, portanto, o número de supressões - se a supressão local for aplicada como um passo adicional.

Aplicamos a supressão local para alcançar o 2-anonimato. A escolha do limiar baixo de dois é baseada no baixo risco geral de reidentificação, devido ao alto peso amostral e a liberação como dados de pesquisa.

Pesos de amostra altos significam, um baixo nível de risco de reidentificação. Alcançar o 2-anonimato é o mesmo que remover registros da amostra. Isto leva a 42 supressões na variável HHSIZE e 4 supressões na variável REGION.

Como explicado anteriormente, a supressão do valor da variável HHSIZE não leva a uma supressão real desta informação. Portanto, refizemos a supressão local, mas desta vez dizemos à sdcMicro para, se possível, não suprimir HHSIZE, mas uma das outras variáveis. Na sdcMicro, é possível dizer ao algoritmo quais variáveis são importantes e quais são menos importantes, para fazer pequenas mudanças.

Para evitar que o HHSIZE seja suprimido, definimos a importância do HHSIZE nos vetores de importância para os mais altos (isto é, 1).

A seguir mostramos como aplicar a supressão local e colocar a importância na variável HHSIZE.

```
# Local suppression

> sdcHH <- localSuppression(sdcHH, k=2, importance = NULL) # no importance vector

> print(sdcHH, "ls")

## Local Suppression:

##     KeyVar | Suppressions (#) | Suppressions (%)

##     URBRUR |               0 |           0.000

##     REGION |               4 |           0.203

##     HHSIZE |              37 |           1.877

##  OWNAGLAND |               0 |           0.000

##      RELIG |               0 |           0.000


> sdcHH <- undolast(sdcHH)
```

```
> sdcHH <- localSuppression(sdcHH, k=2, importance = c(3, 2, 1, 5, 5)) # importance on
                HHSIZE (1), REGION (2) and URBRUR (3)


> print(sdcHH, "ls")

## Local Suppression:

##     KeyVar | Suppressions (#) | Suppressions (%)

##     URBRUR |               6 |           0.304

##     REGION |               1 |           0.051

##     HHSIZE |               1 |           0.051

##  OWNAGLAND |              43 |           2.182

##      RELIG |              16 |           0.812
```

A variável REGION é o tipo de variável que também não deve ter nenhuma supressão. Também definimos a importância de REGION para 2 e a importância de RURURB para 3. Isto leva a uma ordem das variáveis a serem consideradas para supressão pelo algoritmo. Em vez de 42 supressões na variável HHSIZE, isto leva a um valor suprimido na variável HHSIZE, e a 6, 1, 48 e 16 supressões, respectivamente, para as variáveis URBRUR, REGION, OWNAGLAND e RELIG (que definimos como menos importantes).

A importância é claramente refletida no número de supressões. O número total de supressões é maior do que sem um vetor de importância (71 vs. 46), mas o 2-anonimato é alcançado no conjunto de dados com menos supressões nas variáveis HHSIZE e REGION. Removemos a única família com o valor suprimido de HHSIZE (13), para proteger esta família.

A função undolast() restaura o objeto sdcMicro de volta ao estado anterior (isto é, antes de

aplicarmos a supressão local), o que nos permite executar novamente o mesmo comando, mas desta vez com um conjunto vetorial de importância.

A função undolast() só pode ser usada para retroceder um passo.

As variáveis ROOF, TOILET, WATER, ELECTCON, FUELCOOK, OWNMOTORCYCLE, CAR, TV e LIVESTOCK não são variáveis sensíveis e não foram selecionadas como quase-identificadores, porque presumimos que não existem fontes de dados externas contendo estas informações que poderiam ser usadas para vinculação. No entanto, os valores podem ser facilmente observados ou conhecidos pelos vizinhos e, portanto, são importantes, conjuntamente com outras variáveis, para o cenário de reconhecimento espontâneo e para o cenário do vizinho intrometido.

Desta forma, para tornar estas variáveis anônimas, queremos introduzir nelas um baixo nível de incerteza. Portanto, decidimos utilizar PRAM invariantes para as variáveis ROOF, TOILET, WATER, ELECTCON, FUELCOOK, OWNMOTORCYCLE, CAR, TV e LIVESTOCK, onde tratamos LIVESTOCK como uma variável semicontínua devido ao baixo número de valores diferentes.

A seguir ilustramos como aplicar a PRAM:

Pram

```
> set.seed(12345)

> sdcHH <- pram(sdcHH, strata_variables = "REGION", pd = 0.8)

## Number of changed observations:

## - - - - - - - - - - -

## ROOF != ROOF_pram : 98 (4.97%)

## TOILET != TOILET_pram : 151 (7.66%)
```

WATER != WATER_pram : 167 (8.47%)

ELECTCON != ELECTCON_pram : 90 (4.57%)

FUELCOOK != FUELCOOK_pram : 113 (5.73%)

OWNMOTORCYCLE != OWNMOTORCYCLE_pram : 41 (2.08%)

CAR != CAR_pram : 172 (8.73%)

TV != TV_pram : 137 (6.95%)

LIVESTOCK != LIVESTOCK_pram : 149 (7.56%)

Escolhemos o parâmetro pd de pram(), o limite inferior para a probabilidade de que um valor não seja alterado, para ser relativamente alto em 0,8. Podemos escolher um valor alto, porque as variáveis em si não são sensíveis e queremos apenas introduzir um baixo nível de mudanças para minimizar a perda de utilidade. Como a distribuição de muitas das variáveis escolhidas para PRAM depende de REGION, decidimos usar a variável REGION como uma variável de estrato. Desta forma, a matriz de transição é computada para cada região separadamente.

Como a PRAM é um método probabilístico, definimos uma semente para o gerador de números aleatórios antes de aplicar a PRAM para garantir a reprodutibilidade dos resultados.

A PRAM mudou os valores dentro das variáveis de acordo com as matrizes de transição invariantes. Como usamos o método PRAM invariante, as frequências univariadas absolutas permanecem inalteradas. Este não é o caso para as frequências multivariadas.

No Passo 10a comparamos as mudanças nas frequências multivariadas para as variáveis PRAMmed: Selecionamos variáveis de receitas e despesas e a variável LANDSIZEHA como quase-identificadores numéricos, como discutido no Passo 4.

Na Etapa 5, identificamos variáveis de grande interesse para os usuários de nossos dados. Muitos usuários utilizam os dados para medir a desigualdade e os padrões de gastos.

Com base na avaliação de risco, no Passo 6a, decidimos desidentificar a variável LANDSIZEHA, pela codificação superior no valor 40 e arredondar valores menores que 1 para 1 dígito, e valores maiores que 1 para zero dígitos.

O arredondamento dos valores impede a correspondência exata com o registro cadastral disponível. Além disso, agrupamos os valores entre 5 e 40 nos grupos 5 - 19 e 20 – 39 e removemos todos os valores extremos através da codificação superior dos valores.

A seguir mostramos como seguir esses passos em R:

```
# Rounding values of LANDSIZEHA to 1 digit for plots smaller than 1 and

# to 0 digits for plots larger than 1

> sdcHH@manipNumVars$LANDSIZEHA[sdcHH@manipNumVars$LANDSIZEHA <= 1

      & !is.na(sdcHH@manipNumVars$LANDSIZEHA)] <
      round(sdcHH@manipNumVars$LANDSIZEHA[sdcHH@manipNumVars$LANDSIZEH
A     <= 1 & !is.na(sdcHH@manipNumVars$LANDSIZEHA)],digits = 1)

      sdcHH@manipNumVars$LANDSIZEHA[sdcHH@manipNumVars$LANDSIZEHA > 1 &

      !is.na(sdcHH@manipNumVars$LANDSIZEHA)] <
      round(sdcHH@manipNumVars$LANDSIZEHA[sdcHH@manipNumVars$LANDSIZEH
A     > 1 & !is.na(sdcHH@manipNumVars$LANDSIZEHA)], digits = 0)


# Grouping values of LANDSIZEHA into intervals 5-19, 20-39

> sdcHH@manipNumVars$LANDSIZEHA[sdcHH@manipNumVars$LANDSIZEHA >= 5 &
      sdcHH@manipNumVars$LANDSIZEHA < 20 &
      is.na(sdcHH@manipNumVars$LANDSIZEHA)] <- 13

> sdcHH@manipNumVars$LANDSIZEHA[sdcHH@manipNumVars$LANDSIZEHA >= 20 &
```

```
sdcHH@manipNumVars$LANDSIZEHA < 40 &
is.na(sdcHH@manipNumVars$LANDSIZEHA)] <- 30
```

Topcoding values of LANDSIZEHA larger than 40 (also recomputes risk after manual changes)

```
> sdcHH <- topBotCoding(sdcHH, value = 40, replacement = 40, kind = 'top', column =
      'LANDSIZEHA')
```

Results for LANDSIZEHA

```
> table(sdcHH@manipNumVars$LANDSIZEHA)

##   0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9   1   2   3   4  13  30  40

## 188 109  55  30  24  65  22   7  31  16 154 258  53  60 113  18  25
```

Após estas etapas, nenhuma família tem um tamanho de lote único e o número de famílias na amostra com o mesmo tamanho de lote foi aumentado para, pelo menos 7, e, pela remoção de valores extremos, reduzimos o risco de reconhecimento espontâneo, conforme discutido na etapa 6.

Para as variáveis de despesas e receitas, comparamos, com base nos dados reais do estudo de caso, vários métodos.

Como mencionado anteriormente, o principal uso dos dados é calcular medidas de

desigualdade, tais como o coeficiente de GINI.[7]

A recodificação destas variáveis em percentis cria dificuldades no cálculo destas medidas ou altera estas medidas em grande medida e, portanto, não é um método adequado.

Muitas vezes as variáveis de receita e despesa que são liberadas em um conjunto de dados de pesquisa são desidentificados por uma codificação superior. Isto protege os valores extremos superiores, que são os valores que estão em maior risco. A codificação superior (e inferior), entretanto, destrói as informações de desigualdade nos dados, removendo as rendas altas (e baixas). Para não alterar o propósito da coleta de dados inicial, decidimos utilizar a adição de ruído, como método mais adequado.

Para levar em conta o maior risco de valores extremos, adicionamos um nível de ruído maior ainda a esses. Antes de aplicar métodos probabilísticos como a adição de ruído, definimos uma semente para o gerador de números aleatórios. Isto nos permite reproduzir os resultados.

A seguir mostramos como realizar a adição de ruído:

```
# Add noise to income and expenditure variables by category

# Anonymize components

> compExp <- c("TFOODEXP", "TALCHEXP", "TCLTHEXP", "THOUSEXP", "TFURNEXP",
        "THLTHEXP", "TTRANSEXP", "TCOMMEXP", "TRECEXP", "TEDUEXP",
        "TRESHOTEXP", "TMISCEXP")


# Add noise to expenditure variables
```

[7] Explicação sobre GINI em <https://www.ipea.gov.br/desafios/index.php?option=com_content&id=2048:catid=28>

```
> set.seed(123)

> sdcHH <- addNoise(noise = 0.01, obj = sdcHH, variables = compExp, method = "additive")


# Add noise to outliers

> sdcHH <- addNoise(noise = 0.05, obj = sdcHH, variables = compExp, method = "outdect")


# Sum over expenditure categories to obtain consistent totals

> sdcHH@manipNumVars[,'TANHHEXP'] <- rowSums(sdcHH@manipNumVars[,compExp])

> compInc <- c('INCRMT', 'INCWAGE', 'INCFARMBSN', 'INCNFARMBSN', 'INCRENT',
        'INCFIN', 'INCPENSN', 'INCOTHER')


# Add noise to income variables

> sdcHH <- addNoise(noise = 0.01, obj = sdcHH, variables = compInc, method = "additive")


# Add noise to outliers

> sdcHH <- addNoise(noise = 0.05, obj = sdcHH, variables = compInc, method = "outdect")


# Sum over income categories to obtain consistent totals

>sdcHH@manipNumVars[,'INCTOTGROSSHH'] <-
        rowSums(sdcHH@manipNumVars[,compInc])
```

recalculate risks after manually changing values in sdcMicro object

> calcRisks(sdcHH)

A adição de ruído pode levar a uma transformação da forma da distribuição. Dependendo da magnitude do ruído, os valores de renda também podem se tornar negativos. Uma maneira de resolver isto seria cortar os valores abaixo de zero e fixá-los em zero. Isto destruiria, entretanto, as propriedades conservadas pela adição de ruído (entre outros, o valor da média esperada) e optamos por manter os valores negativos.

Como mencionado anteriormente, os agregados de receitas e despesas são as somas dos componentes. A adição de ruído a cada um dos componentes pode levar à violação desta condição. Portanto, uma solução é adicionar ruído aos agregados e remover os componentes.

Preferimos manter os componentes nos dados e aplicar a adição de ruído a cada componente separadamente. Isto permite aplicar um nível de ruído menor do que quando se aplica ruído somente aos agregados. Um nível de ruído de 0,01 parece ser suficiente com ruído extra de 0,05 adicionado aos valores extremos. Após adicionar o ruído aos componentes, recalculamos os agregados como a soma dos componentes perturbados.

3.9 – Etapa 9a: Remensuração do risco

Para as variáveis categóricas, concluímos que obtivemos 2-anonimato nos dados com supressão local. Apenas 104 domicílios, ou cerca de 5% do número total, violam o 3-anonimato. A Tabela 34 dá uma visão geral dessas medidas de risco. O risco global é reduzido para 0,02% (número esperado de reidentificações 0,36), o que é extremamente baixo. Portanto, concluímos que, com base nas variáveis categóricas, os dados foram suficientemente desidentificados.

Nenhum agregado familiar tem um risco de reidentificação maior que 0,01 (1%). Ao remover os domicílios com valores raros (ou aberrantes) da variável HHSIZE, reduzimos o risco de reconhecimento espontâneo desses domicílios. Este raciocínio também pode ser aplicado ao resultado do risco de recodificação da variável LANDSIZEHA e a pós aleatorização das variáveis identificadas como importantes no cenário do vizinho intrometido. Desta forma, um intruso não consegue saber, com certeza, se uma família que ele reconhece nos dados é a família correta, devido ao ruído.

Estas medidas se referem apenas às variáveis categóricas. Para avaliar o risco das variáveis contínuas, poderíamos usar uma medida de intervalo ou um algoritmo de vizinho mais próximo. Escolhemos usar uma medida de intervalo, uma vez que a vinculação exata do valor não é nossa maior preocupação, com base nos cenários assumidos e nas fontes de dados externas. Em vez disso, conjuntos de dados com valores semelhantes, mas não exatamente os mesmos valores, poderiam ser usados para a vinculação. Aqui, a principal preocupação é que os valores estejam suficientemente distantes dos valores originais, que são medidos com uma medida de intervalo.

A seguir, mostramos como avaliar a medida de intervalo para cada uma das variáveis de despesa, indicadas no vetor compExp:

```
> dRisk(sdcHH@origData[,compExp], xm = sdcHH@manipNumVars[,compExp], k = 0.01)

##   [1] 0.0608828
```

```
> dRisk(sdcHH@origData[,compExp], xm = sdcHH@manipNumVars[,compExp], k = 0.02)

##   [1] 0.9025875
```

```
> dRisk(sdcHH@origData[,compExp], xm = sdcHH@manipNumVars[,compExp], k = 0.05)
```

[1] 1

Os diferentes valores do parâmetro *k* na função dRisk() definem o tamanho do intervalo em torno do valor original. Quanto maior *k*, maiores os intervalos, maior a probabilidade de que um valor perturbado esteja no intervalo em torno do valor original e maior a medida de risco. O resultado é satisfatório com intervalos relativamente pequenos (*k* = 0,01), mas não quando se aumenta o tamanho dos intervalos.

Em nosso caso, *k* = 0,01, é suficientemente grande, já que estamos olhando para os componentes, não para os agregados.

Temos que prestar atenção especial aos valores extremos. Aqui o valor 0,01 para *k* é muito pequeno para supor que eles estejam protegidos quando fora deste pequeno intervalo. Seria necessário verificar os valores extremos e seus valores perturbados e poderia haver a necessidade de um nível mais alto de perturbação para os valores extremos. Concluímos que, de uma perspectiva de risco e com base na medida do intervalo, os níveis de ruído escolhidos são aceitáveis.

Na próxima etapa, analisaremos o impacto da adição de ruído sobre a utilidade dos dados.

3.10 – Etapa 10a: Remensuração da utilidade

Nenhuma das variáveis foi recodificada e o nível original de detalhe nos dados é mantido, exceto para a variável LANDSIZEHA.

Como descrito na etapa 8a, a supressão local removeu apenas alguns valores nas outras variáveis, o que não reduziu muito a validade dos dados. As distribuições univariadas de frequência das variáveis ROOF, TOILET, WATER, ELECTCON, FUELCOOK, OWNMOTORCYCLE, CAR, TV e LIVESTOCK não mudaram em grande parte, por definição do método PRAM invariante. Entretanto, alguns valores foram trocados entre os domicílios. Isto se torna aparente quando se observa as frequências multivariadas de WATER com a variável URBRUR.

As frequências multivariadas do PRAMmed com a variável URBRUR poderiam ser de interesse para os usuários, mas estas não são preservadas.

Como aplicamos PRAM dentro das regiões, as frequências multivariadas das variáveis PRAMmed com a variável REGION são preservadas. Para concisão, nos restringiremos à análise das variáveis de gastos. A análise das variáveis de renda pode ser feita da mesma forma e leva a resultados semelhantes.

Analisamos o efeito da desidentificação em alguns indicadores de padrão, conforme discutido na etapa 5.

Primeiramente, as estimativas pontuais e o intervalo de confiança inicial do coeficiente GINI para a soma dos componentes da despesa. O cálculo do coeficiente GINI e o intervalo de confiança são baseados nos valores positivos das despesas.[8] Enquanto para os dados originais obtivemos coeficiente GINI de 0,510 (0,476-0,539), para os dados desidentificados obtivemos 0,508 (0,476-0,538).

Como pode-se observar, as mudanças no coeficiente GINI foram muito pequenas, e estatisticamente insignificantes. Além do coeficiente GINI, comparamos as despesas médias mensais (MME) e a renda média mensal (MMI) para a população rural, urbana e total.

Para MME, nos dados originais encontramos 400,5, 457,3 e 412,6, enquanto para os dados desidentificados encontramos 398,5, 459,9 e 412,6, respectivamente para MME rural, urbano e total.

Já, para MMI, encontramos nos dados originais os valores 397,1, 747,6 e 472,1, enquanto que para os dados desidentificados encontramos 402,2, 767,8 e 478,5 respectivamente para MMI rural, urbano e total.

Observamos que os níveis de ruído escolhidos adicionaram apenas pequenas distorções à MME e mudanças ligeiramente maiores à MMI.

A desidentificação para a criação de um arquivo de pesquisa levará inevitavelmente a algum

[8] A explicação detalhada dessa medida e de outras usadas como indicadores de padrão nessa seção estão fora do escopo desse manual.

grau de perda de utilidade. É importante descrever esta perda no relatório externo, para que os usuários estejam cientes das mudanças nos dados.

O próximo passo é fundir as variáveis domiciliares tratadas com as variáveis individuais não tratadas. Assim, realizamos a desidentificação, agora, das variáveis de nível individual. Isto também inclui a seleção das variáveis utilizadas na desidentificação das variáveis de nível individual.

A seguir mostramos os passos para fundir estes arquivos.

```
### Select variables (individual level)

# Key variables (individual level)

> selectedKeyVarsIND = c('GENDER', 'REL', 'MARITAL', 'AGEYRS', 'EDUCY', 'ATSCHOOL',
        'INDUSTRY1') # list of selected key variables


# Sample weight (WGTHH, individual weight)

> selectedWeightVarIND = c('WGTHH')


# Household ID

> selectedHouseholdID = c('IDH')


# No strata

# Recombining anonymized HH datasets and individual level variables

> indVars <- c("IDH", "IDP", selectedKeyVarsIND, "WGTHH") # HID and all non HH variables
```

```
> fileInd <- file[indVars] # subset of file without HHVars


> HHmanip <- extractManipData(sdcHH) # manipulated variables HH

> HHmanip <- HHmanip[HHmanip[,'IDH'] != 1782,]


> fileCombined <- merge(HHmanip, fileInd, by.x= c('IDH'))

> fileCombined <- fileCombined[order(fileCombined[,'IDH'],fileCombined[,'IDP']),]


# SDC objects with all variables and treated HH vars for

# anonymization of individual level variables

> sdcCombined <- createSdcObj(dat = fileCombined, keyVars = selectedKeyVarsIND,
      weightVar = selectedWeightVarIND, hhId = selectedHouseholdID)
```

Criamos o objeto sdcMicro para a desidentificação das variáveis individuais da mesma forma que para a variável de nível domiciliar. Posteriormente, repetimos as etapas 6-10 para as variáveis de nível individual.

3.11 – Etapa 6b: Avaliação do risco de divulgação (nível individual)

Todas as variáveis-chave em nível individual são categóricas. Portanto, podemos usar o k-anonimato e as medidas de risco individual e global.

O risco hierárquico é agora de interesse, dada a estrutura domiciliar no arquivo do conjunto de dados sdcCombined, que inclui tanto as variáveis de nível doméstico quanto as de nível

individual. O número de indivíduos (absolutos e relativos) que violam o k-anonimato nos limiares 2, 3 e 5 são, respectivamente, de 998 (9,91%), 1.384 (13,75%) e 2.194 (21,79%).

As medidas de risco global podem ser encontradas usando R, como a seguir:

```
> print(sdcCombined, 'risk')

## Risk measures:

##

## Number of observations with higher risk than the main part of the data: 0

## Expected number of re-identifications: 23.98 (0.24 %)

##

## Information on hierarchical risk:

## Expected number of re-identifications: 127.12 (1.26 %)
```

O risco global é de 0,24%, o que corresponde a 24 reidentificações esperadas. Considerando a estrutura hierárquica, isto sobe para 1,26%, ou 127 reidentificações esperadas. As medidas de risco global são baixas em comparação com o número de violadores do k-anonimato, devido ao baixo peso de amostragem.

O alto número de violadores do k-anonimato ocorre, principalmente, em razão da variável de idade muito detalhada.

As medidas de risco baseiam-se apenas nas variáveis de nível individual, uma vez que assumimos que as variáveis de nível individual e domiciliar não são usadas simultaneamente por um intruso. Se considerarmos um cenário de intruso onde estas variáveis são usadas simultaneamente por um intruso para reidentificar indivíduos, as variáveis de nível doméstico

também devem ser levadas em consideração aqui. Isto resultaria em um alto número de variáveis-chave.

3.12 – Etapa 7b: Avaliação das medidas de utilidade (nível individual)

Avaliamos as medidas de utilidade, selecionadas na etapa 5, além de algumas medidas de utilidade geral. Os valores calculados, a partir dos dados brutos, serão apresentados na etapa 10b para permitir a comparação direta com os valores calculados, a partir dos dados desidentificados.

3.13 – Etapa 8b: Escolha e aplicação dos métodos SDC (nível individual)

Usamos a mesma abordagem para a desidentificação das variáveis-chave categóricas de nível individual, como para as variáveis categóricas de nível domiciliar, descritas anteriormente. Desta forma, primeiro usamos a recodificação global para limitar o número necessário de supressões, depois aplicamos supressões locais e, finalmente, quando necessário, usamos métodos perturbativos.

A variável AGEYRS, relativa à idade, em anos, tem muitos valores diferentes - idade, em meses, para crianças de 0 a 1 ano e idade, em anos, para indivíduos acima de 1 ano. Este nível de detalhe leva a um alto risco de reidentificação, em razão dos conjuntos de dados externos com a idade exata, bem como o conhecimento da idade exata de parentes próximos. Por isso, temos que reduzir o nível de detalhe nas variáveis de idade, recodificando esses valores.

Então, primeiro, recodificamos os valores de 15 a 65 anos, em intervalos de dez anos - como alguns indicadores relacionados à educação são computados a partir do conjunto de dados da pesquisa, nossa primeira abordagem não é recodificar a faixa etária de 0 a 15 anos. Para crianças com menos de 1 ano de idade, reduzimos os detalhes, recodificando para 0 anos, assim como protegemos os indivíduos com altos (raros) valores de idade, limitando a idade máxima em 65 anos.

A seguir, mostramos como obter essas recodificações no R:

```
# Recoding age and top coding age (top code 65), below that 10 year age
# groups, children aged under 1 are recoded 0 (previously in months)
> sdcCombined@manipKeyVars$AGEYRS[sdcCombined@manipKeyVars$AGEYRS >= 0 &
        sdcCombined@manipKeyVars$AGEYRS < 1] <- 0
> sdcCombined@manipKeyVars$AGEYRS[sdcCombined@manipKeyVars$AGEYRS >= 15
&
        sdcCombined@manipKeyVars$AGEYRS < 25] <- 20
...
> sdcCombined@manipKeyVars$AGEYRS[sdcCombined@manipKeyVars$AGEYRS >= 55
&
        sdcCombined@manipKeyVars$AGEYRS < 66] <- 60

# topBotCoding also recalculates risk based on manual recoding above
> sdcCombined <- topBotCoding(obj = sdcCombined, value = 65, replacement = 65, kind =
        'top', column = 'AGEYRS')
> table(sdcCombined@manipKeyVars$AGEYRS) # check results
##   0   1   2   3   4   5   6   7   8   9  10  11  12  13  14
## 311 367 340 332 260 334 344 297 344 281 336 297 326 299 263
##  20  30  40  50  60  65
```

1847 1220 889 554 314 325

Estas recodificações já reduzem o risco para 531 indivíduos que violam o 3-anonimato.Poderíamos recodificar os valores de idade na faixa etária inferior, de acordo com as categorias etárias que os usuários exigem - por exemplo, 8-11 para educação. Isto reduziria a utilidade dos dados para alguns usuários. Portanto, decidimos olhar primeiro para o número de supressões necessárias na supressão local.

A seguir, demonstramos como se pode experimentar a supressão local para encontrar a melhor opção:

```
# Copy of sdcMicro object to later undo steps

> sdcCopy <- sdcCombined


# Importance vectors for local suppression (depending on utility measures)

> impVec1 <- NULL # for optimal suppression

> impVec2 <- rep(length(selectedKeyVarsIND), length(selectedKeyVarsIND))

> impVec2[match('AGEYRS', selectedKeyVarsIND)] <- 1 # AGEYRS

> impVec2[match('GENDER', selectedKeyVarsIND)] <- 2 # GENDER


# Local suppression without importance vector

> sdcCombined <- localSuppression(sdcCombined, k = 2, importance = impVec1)
```

Number of suppressions per variable

> print(sdcCombined, "ls")

Local Suppression:

KeyVar | Suppressions (#) | Suppressions (%)

GENDER | 0 | 0.000

REL | 34 | 0.338

MARITAL | 0 | 0.000

AGEYRS | 195 | 1.937

EDUCY | 0 | 0.000

EDYRSCURRAT | 3 | 0.030

ATSCHOOL | 0 | 0.000

INDUSTRY1 | 21 | 0.209

Number of suppressions per variable for each value of AGEYRS

>table(sdcCopy@manipKeyVars$AGEYRS) - table(sdcCombined@manipKeyVars$AGEYRS)

0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 20 30 40 50 60 65

0 0 0 0 0 0 2 0 2 1 0 1 4 1 5 25 53 37 36 15 13

```
# Undo local suppression

> sdcCombined <- undolast(sdcCombined)


# Local suppression with importance vector on AGEYRS and GENDER

> sdcCombined <- localSuppression(sdcCombined, k = 2, importance = impVec2)


# Number of suppressions per variable

> print(sdcCombined, "ls")

## Local Suppression:

##       KeyVar | Suppressions (#) | Suppressions (%)

##       GENDER |               0 |           0.000

##          REL |             323 |           3.208

##      MARITAL |               0 |           0.000

##       AGEYRS |               0 |           0.000

##        EDUCY |               0 |           0.000

##  EDYRSCURRAT |               0 |           0.000

##     ATSCHOOL |               0 |           0.000

##    INDUSTRY1 |               0 |           0.000


# Number of suppressions for each value of the variable AGEYRS
```

>table(sdcCopy@manipKeyVars$AGEYRS) - table(sdcCombined@manipKeyVars$AGEYRS)

0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 20 30 40 50 60 65

0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

Usamos a supressão local para alcançar 3-anonimato. Na primeira tentativa, não especificamos nenhum vetor de importância, o que levou a muitas supressões na variável AGEYRS. Isto é indesejável de um ponto de vista de utilidade. Portanto, decidimos especificar um vetor de importância para evitar supressões na variável AGEYRS.

A supressão da variável GENDER também é indesejável do ponto de vista da utilidade. A variável GENDER é um tipo de variável que não deve ter supressões. Nós definimos GENDER como a variável com a segunda maior importância. Depois de especificar o vetor de importância, para evitar supressões da variável idade, não há supressões de idade.

O número total de supressões nas outras variáveis aumentou, entretanto, de 253 para 323 por causa do vetor de importância. Isto é de se esperar, porque o algoritmo sem o vetor de importância minimiza o número total de supressões, através da primeira supressão de valores em variáveis com muitas categorias - neste caso, a idade e o sexo.

Depois de especificar um vetor de importância, a variável REL teve muitas supressões. Nós escolhemos esta segunda opção.

3.14 – Etapa 9b: Remensuração do Risco (nível individual)

Reavaliamos as medidas de risco selecionadas na etapa 6b. Encontramos que 197 (1,96%) e 518 (5,15%) indivíduos violavam, respectivamente, o 2- e 3-anonimato. A supressão local, não surpreendentemente, reduziu o número de indivíduos que violam o 2-anonimato para 0. O risco global hierárquico foi reduzido para 0,11%, o que corresponde a 11,3 reidentificações esperadas. O maior risco individual de reidentificação hierárquica é de 1,21%. Estes níveis

de risco parecem aceitáveis para um arquivo de pesquisa.

3.15 – Etapa 10b: Remensuração da Utilidade (nível individual)

Selecionamos duas medidas de utilidade para as variáveis individuais: Matrícula no ensino primário e secundário, ambas também por gênero. Estas duas medidas são sensíveis às mudanças nas variáveis gênero (GENDER), idade (AGEYRS) e educação (EDUCY e EDYRSATCURR), e, portanto, dão uma boa visão geral do impacto da desidentificação. Antes, para os dados originais, encontramos para a educação primária: Entre meninos, 74,2%; meninas, 70,9% e, total, 72,6%. Já, para a educação secundária: Entre meninos, 44,8%, meninas, 39,1%, e, total, 42,0%. Nessa mesma ordem, encontramos para os dados desidentificados, 74,2%, 70,9% e 72,6% - para a educação primária - e, 44,8%, 39,1%, e, 42,0%, para a educação secundária.

Como pode-se observar, a desidentificação não alterou os resultados.

3.16 – Etapa 11: Auditoria e relatórios

Numa primeira etapa de auditoria, verificamos se os dados permitem a reprodução dos números publicados do conjunto de dados original e, se as relações entre as variáveis e outras características dos dados, são preservadas no processo de desidentificação.

Em resumo, verificamos se o conjunto de dados é válido para fins analíticos.

Na segunda etapa, exploramos as características dos dados e as relações entre as variáveis. Estas características e relações de dados foram preservadas principalmente porque as levamos em consideração ao escolher os métodos de desidentificação apropriados.

As variáveis TANHHEXP e INCTOTGROSSHH são as somas dos componentes individuais, pois adicionamos ruído aos componentes e reconstruímos os agregados através da soma sobre os componentes.

Inicialmente, as variáveis de renda eram todas positivas. Esta característica foi violada, como resultado da adição de ruído. Como os valores da variável AGEYRS não foram perturbados, mas apenas recodificados e suprimidos, não introduzimos combinações improváveis, tais como um indivíduo de 60 anos matriculado no ensino primário. Além disso, ao separar o processo de desidentificação em duas partes, uma para variáveis de nível domiciliar e outra para variáveis de nível individual, os valores das variáveis medidas no nível domiciliar concordam para todos os membros de cada domicílio.

Se faz necessária a elaboração de dois relatórios, interno e externo, sobre a desidentificação do conjunto de dados. O relatório interno inclui os métodos utilizados, o risco antes e depois da desidentificação, assim como as razões para os métodos selecionados e seus parâmetros. Já, o relatório externo enfoca as mudanças nos dados e a perda de utilidade. O foco aqui deve ser o número de supressões, bem como os métodos perturbadores (PRAM). Isto está descrito nas etapas anteriores. Dependendo dos usuários e leitores dos relatórios, o conteúdo pode ser diferente.

A função report() utiliza os dados disponíveis no objeto sdcMicro para gerar relatórios como mostrado a seguir.

```
# Create reports with sdcMicro report() function

> report(sdcHH, internal = F) # external (brief) report

> report(sdcHH, internal = T) # internal (extended) report
```

3.17 – Etapa 12: Liberação de dados

A etapa final é a liberação do conjunto de dados anonimizados junto com o relatório externo. A seguir mostramos como coletar os dados do objeto sdcMicro com a função extractManipData().

Antes de liberar o arquivo, adicionamos uma identificação individual ao arquivo (número de

linha no domicílio). Depois, exportamos o conjunto de dados anonimizados como arquivo STATA.

```
# Anonymized dataset

# Household variables and individual variables

# extracts all variables, not just the manipulated ones

> dataAnon <- extractManipData(sdcCombined, ignoreKeyVars = F, ignorePramVars = F,
      ignoreNumVars = F, ignoreStrataVar = F)


# Create STATA file

> write.dta(dataframe = dataAnon, file= 'Case1DataAnon.dta', convert.dates=TRUE)
```

4 – Considerações Finais

Este manual buscou propor o uso de uma ferramenta, disponibilizada gratuitamente na internet – o programa R e o pacote sdcMicro, para a realização do processo de anonimização, mostrando o passo a passo, através de um caso concreto hipotético, utilizado como exemplo na dinâmica.

Como mencionado na introdução, a anonimização deve fazer parte do planejamento do pesquisador que pretende divulgar sua pesquisa e esta ferramenta gratuita, se utilizada adequadamente, permite atender aos parâmetros legais nacionais vigentes sobre anonimização reversível, pseudonimização, ou até uma anonimização irreversível.

Existem outros programas/pacotes e outras possibilidades de atender aos ditames legais,

mas, especialmente, esclarecemos que cabe ao pesquisador ou qualquer pessoa/operador que assuma a obrigação de trabalhar com dados pessoais de terceiros, que antecipadamente considere os riscos de prejuízos para estes terceiros no caso de vazamentos, na possibilidade de cruzamentos desses dados com informações disponíveis na internet ou em outros conjuntos de dados.

Referências

BRASIL. Lei nº13.709/2018. Disponível em <https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm>

International Household Survey Network (IHSN). Disponível em <http://ihsn.org/sites/default/files/resources/case_studies_code_and_data.zip>

R Project. Disponível em <http://cran.r-project.org/>.

RStudio. Disponível em <http://www.rstudio.com>

Templ, Matthias (2017). Statistical Disclosure Control for Microdata: Methods and Applications in R. Springer. ISBN 978-3-319-50272-4