São José dos Campos - SP
LV SBPO
LV Simpósio Brasileiro de Pesquisa Operacional
São José dos Campos, SP - 6 a 9 de novembro de 2023

HEURÍSTICA BRKGA APLICADA À RESOLUÇÃO DO PROBLEMA DE MICROAGREGAÇÃO MULTIVARIADA

Augusto Cesar Fadel
Universidade Federal Fluminense - UFF
Avenida General Milton Tavares de Souza, s/n, São Domingos, Niterói - RJ
augustofadel@gmail.com

Luiz Satoru Ochi
Universidade Federal Fluminense - UFF
Avenida General Milton Tavares de Souza, s/n, São Domingos, Niterói - RJ
satoru@ic.uff.br

José André de Moura Brito
Escola Nacional de Ciências Estatísticas - ENCE/IBGE
Rua André Cavalcanti, 106, sala 503-A, Bairro de Fátima, Rio de Janeiro - RJ
jose.m.brito@ibge.gov.br

RESUMO

No mundo atual, instituições públicas e privadas coletam e armazenam grandes volumes de dados, mas apenas uma pequena parte é disponibilizada à sociedade, geralmente em função de questões éticas e legais relacionadas à privacidade dos dados pessoais. O compartilhamento de dados entre instituições pode melhorar a eficiência do uso de recursos e da formulação de políticas públicas, mas a garantia da privacidade é fundamental. Neste contexto, técnicas de controle estatístico de confidencialidade e tecnologias de aprimoramento da privacidade oferecem soluções para viabilizar o uso destes dados de forma segura, promovendo transparência e aumentando a confiança da sociedade. A microagregação é uma técnica capaz de reduzir os riscos à privacidade, mantendo, porém, a utilidade dos dados. O presente trabalho apresenta os resultados de um método efetivo no controle destes objetivos, superando outros métodos da literatura.

PALAVRAS CHAVE. Metaheurísticas, Microagregação, BRKGA.

MH, OC

ABSTRACT

Nowadays, public and private institutions collect and store large volumes of data, but only a small part is made available to society, usually due to ethical and legal issues related to data privacy. Sharing data between institutions can improve the use of public resources and more effective policy making. In this context, statistical disclosure control techniques and privacy-enhancing technologies offer solutions to enable the safe use of personal data, promoting transparency and increasing trust and acceptance by society. Microaggregation is a technique capable of reducing privacy risks while maintaining data utility. The present work bring the results of an effective method in controlling these objectives, surpassing other methods in the literature.

KEYWORDS. Metaheuristics. Microaggregation. BRKGA.

MH, OC

DOI: 10.59254/sbpo-2023-174920
http://live-proceedings.pantheonsite.io/p/174920

1. Introdução

Nos dias atuais, instituições públicas e privadas coletam e armazenam volumes cada vez maiores de dados. Entretanto, segundo Domingo-Ferrer e Muralidhar [2015], geralmente apenas uma pequena parte desses dados é disponibilizada à sociedade, tendo como objetivos, por exemplo, a transparência, o planejamento, a identificação de oportunidades de negócios e o bem-estar geral.

As instituições públicas, com destaque para os Institutos Nacionais de Estatística (INE), detêm bases de dados de grande cobertura, por vezes censitária, e de interesse público em sua maioria. Entretanto, aspectos éticos e legais, relacionados à garantia da privacidade de dados pessoais (alguns considerados sensíveis), restringem a divulgação ou disponibilização de tais informações. A privacidade é um direito fundamental, assegurado na Declaração Universal dos Direitos Humanos, cláusula pétrea da Constituição da República Federativa do Brasil de 1988 e objeto central da Lei Geral de Proteção de Dados Pessoais (LGPD). Os INE, ao mesmo tempo que devem resguardar dados pessoais sensíveis, têm um compromisso democrático com a transparência.

Além do compromisso com igualdade de acesso e transparência frente à sociedade, o compartilhamento de dados entre instituições de diferentes esferas de governo, bem como academia e setor privado, tem enorme potencial de contribuir no uso mais eficiente de recursos públicos, seja ao evitar duplicação de esforços na coleta e armazenamento, seja na elaboração, aplicação e acompanhamento de políticas públicas mais efetivas.

No contexto das cidades inteligentes, onde grandes quantidades de dados são coletados de várias fontes para análise e tomada de decisão, a proteção da privacidade tem importância crucial. Cidades inteligentes usam dados de indivíduos, como padrões de transporte, consumo de energia ou atividade de mídia social, para otimizar a alocação de recursos, em sua maioria públicos, melhorar os serviços urbanos possibilitando, consequentemente, uma melhoria na qualidade de vida da população em geral.

Em particular, dados de localização obtidos através de aparelhos de telefonia móvel são de grande valor em estudos de mobilidade urbana e turismo sustentável, por exemplo. Entretanto, despertam preocupação no que diz respeito a uma possível vigilância, gerando possível ”resistência”do usuário quanto ao fornecimento/acesso às suas informações. De forma a mitigar tal resistência e poder utilizar tais informações, as técnicas de controle estatístico de confidencialidade (do inglês, *Statistical Disclosure Control – SDC*) e as tecnologias de aprimoramento da privacidade (do inglês, *Privacy Enhancing Technologies* – PET) são capazes de viabilizar tais estudos garantindo, concomitantemente, a privacidade do indivíduo, conforme discutido em profundidade em Elliot e Domingo-Ferrer [2018] e ONU [2023], por exemplo.

Neste sentido, a microagregação multivariada pode contribuir em diversos aspectos, dentre os quais destacamos três de grande relevância, quais sejam: (i) preservação da privacidade, uma vez que a desidentificação (ou pseudoanonimização), isto é, a remoção de informações de identificação pessoal, não é suficiente para evitar a reidentificação; a aplicação de técnicas como a microagregação permite controlar o risco de reidentificação e proteger a privacidade dos indivíduos; (ii) utilidade dos dados, pois a microagregação permite equilibrar a proteção da privacidade, em termos do risco de reidentificação, e a utilidade dos dados, permitindo análises e tomadas de decisão mais relevantes e assertivas; (iii) confiança e aceitação por parte da sociedade civil: as preocupações com a privacidade podem afetar significativamente a confiança do público e a aceitação de iniciativas pautadas no uso mais intensivo de dados pessoais, como algumas relacionadas a cidades inteligentes. Ao implementar técnicas robustas de proteção de privacidade, como microagregação, o poder público demonstra um compromisso em proteger os dados dos indivíduos e promover a transparência, aumentando a confiança e a aceitação por parte da sociedade.

DOI: 10.59254/sbpo-2023-174920
http://live-proceedings.pantheonsite.io/p/174920

O presente trabalho está organizado em quatro seções, além desta introdução. Na Seção 2 são apresentadas a descrição e formalização do Problema de Microagregação Multivariado (do inglês, *Multivariate MicroAggregation Problem* – MMAP). Na Seção 3 é descrita a metodologia proposta pelos autores para abordá-lo. A Seção 4 apresenta os resultados dos experimentos computacionais realizados e, finalmente, a Seção 5 reúne as principais conclusões obtidas.

2. Descrição do Problema

O Problema de Microagregação Multivariado (do inglês, *Multivariate MicroAggregation Problem* – MMAP) pode ser formalizado como: dado um conjunto de dados T, com n objetos descritos por f atributos quantitativos[1], k grupos, G_i $(i \in \{1, \cdots, k\})$, de tamanho mínimo c (nível de agregação) devem ser construídos, de forma a minimizar, simultaneamente, a perda de informação (do inglês, *Information Loss* - IL) e o risco de reidentificação ou violação de sigilo (do inglês, *Disclosure Risk* - DR) no conjunto resultante T', que, após agregação, continua a ter n objetos e f atributos, porém modificados. Além da restrição de tamanho mínimo dos grupos, todo objeto de T deve ser associado a um e apenas um grupo; o que implica observar as seguintes restrições na construção dos grupos: $\bigcup_{i=1}^{k} G_i = T$, $G_i \cap G_j \neq \emptyset$, $i \neq j$ e $|G_i| \geq c$,$\forall\ i, j \in \{1, \cdots, k\}$.

Duas etapas são necessárias para obter uma solução para este problema, sejam elas: (i) particionamento, onde o conjunto de dados T, composto por n objetos, descritos por f atributos, é particionado em k grupos, de tal maneira que os objetos alocados em um mesmo grupo sejam similares entre si, segundo uma métrica definida (função dos atributos), e que cada grupo tenha, ao menos, c objetos; (ii) agregação, onde um operador de agregação (e.g. média, para atributos numéricos) é utilizado para calcular o centroide de cada grupo. Os valores dos atributos de cada objeto são então substituídos pelos valores correspondentes ao centroide (vetor de médias dos atributos) do grupo ao qual o objeto foi alocado na primeira etapa, produzindo o conjunto T'.

A Figura 1 ilustra o MMAP em um conjunto de dados hipotético. Seja T um conjunto de dados com $n = 9$ objetos e $f = 2$ atributos (idade e renda), representado na Figura 1a.

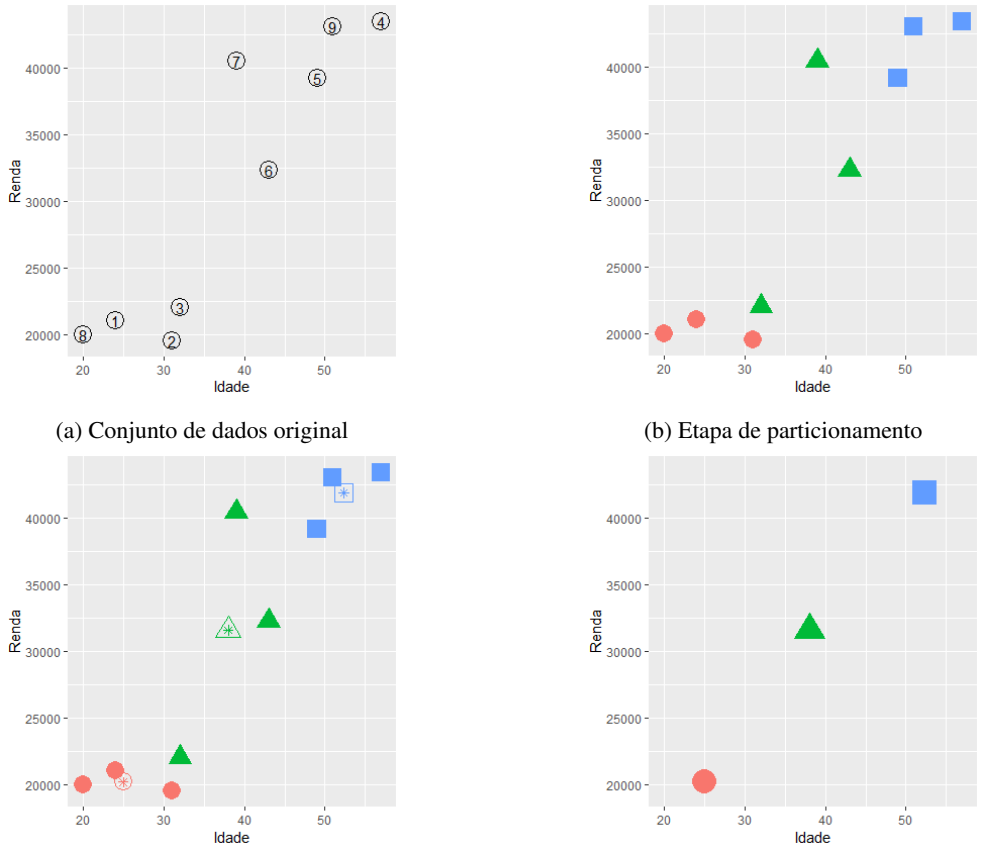
Segundo Domingo-Ferrer e Mateo-Sanz [2002], em uma solução ótima do MMAP o tamanho de cada grupo G_i, doravante definido por n_i, é menor do que $2c$, para todo $i \in \{1, \cdots, k\}$. Dessa forma, para o nível de agregação $c = 3$, o tamanho dos grupos é dado por $3 \leq n_i < 6$, $i = 1, 2, 3$.

A Figura 1b apresenta a etapa de particionamento, na qual é produzida uma partição válida do conjunto de dados. Os objetos alocados aos grupos G_1, G_2 e G_3 são indicados, respectivamente, por círculos, triângulos e quadrados. Na Figura 1c, os asteriscos representam os centroides dos grupos, identificados pelo contorno da respectiva forma geométrica.

Finalmente, a Figura 1d apresenta a etapa de agregação, onde cada objeto tem o valor de seus atributos substituído pelos valores correspondentes ao centroide do grupo. Nessa etapa ocorre, em cada grupo, a perturbação do dado através da representação de c objetos com os mesmos valores de atributos, reduzindo a informação disponível e o risco de violação de sigilo. Quanto maior o nível de agregação, maior é a perda de informação e menor é o risco de violação de sigilo no conjunto de dados resultante T', que será disponibilizado ao usuário final.

[1]Originalmente, o MMAP considerava apenas atributos numéricos contínuos Mortazavi e Jalili [2015]. Embora existam métodos de microagregação para tratar outros tipos de atributos Domingo-Ferrer e Torra [2005], neste trabalho foram considerados apenas atributos quantitativos.

DOI: 10.59254/sbpo-2023-174920
http://live-proceedings.pantheonsite.io/p/174920



(a) Conjunto de dados original

(b) Etapa de particionamento

(c) Determinação dos centroides

(d) Etapa de agregação

Figura 1: MMAP aplicado a um conjunto de dados hipotético

Muitas abordagens já foram propostas na literatura para resolver o MMAP. Domingo-Ferrer e Torra [2005] propuseram o MDAV (Maximum Distance to Average Vector), que, segundo Solanas [2008], é o método de microagregação multivariado mais utilizado e um dos mais eficientes e eficazes em termos da redução de IL. Técnicas de redução de dimensionalidade, como a técnica de análise de componentes principais (Jolliffe [2002]), são aplicadas em métodos como o PPPCA (Projection Pursuit Principal Components Analysis), proposto por Croux e Ruiz-Gazen [2005], e o clustPPPCA (Clustered Projection Pursuit Principal Components Analysis), proposto por Templ [2006]. Solanas et al. [2006] propôs método baseado em um algoritmo genético que apresentou desempenho superior ao método MDAV para conjuntos de dados de até 50 objetos. Apesar da variedade de abordagens, a grande maioria dos métodos disponíveis buscam minimizar apenas a perda de informação (IL), sendo DR controlado indiretamente pelo nível de agregação c e, em geral, desconsiderado na avaliação da qualidade das soluções. Neste trabalho é proposto e comparado método de microagregação multivariado que minimiza diretamente ambos os objetivos.

3. Metodologia

Como apresentado na Seção 2, dois objetivos conflitantes devem ser considerados no MMAP, quais sejam: a perda de informação (IL) e o risco de violação de sigilo (DR). A perda de informação (IL) ocorre na etapa de agregação e é acentuada se, na etapa de particionamento, objetos pouco similares são alocados em um mesmo grupo. Uma solução ótima para o problema de microagregação, no que diz respeito à perda de informação, é aquela na qual a similaridade entre os

DOI: 10.59254/sbpo-2023-174920
http://live-proceedings.pantheonsite.io/p/174920

objetos de cada grupo (homogeneidade interna) é máxima, segundo Fayyoumi e Oommen [2010].

A medida de IL mais utilizada na literatura de Controle Estatístico de Confidencialidade (do inglês, *Statistical Disclosure Control – SDC*) (Domingo-Ferrer et al. [2006], Domingo-Ferrer e Mateo-Sanz [2002], Fayyoumi e Oommen [2010]) é dada pela razão entre a soma do erro quadrático (do inglês, *Sum of Squared Errors* - SSE) e a soma do erro quadrático total (do inglês, *Sum of Squares Total* - SST), conforme descrito na Equação 1, onde x_{jl} corresponde ao valor do atributo l para o objeto j e $\bar{x}_{il}$ é a média do atributo l, considerando os objetos pertencentes ao grupo G_i, para $i \in \{1, \cdots, k\}$, $j \in \{1, \cdots, n\}$ e $l \in \{1, \cdots, f\}$.

$$IL(T, T') = \frac{SSE}{SST} = \frac{\sum_{l=1}^{f} \sum_{i=1}^{k} \sum_{\forall x_{jl} \in G_i} (x_{jl} - \bar{x}_{il})^2}{\sum_{l=1}^{f} \sum_{j=1}^{n} (x_{jl} - \bar{x}_l)^2} \tag{1}$$

Porém, quando apenas IL é adotada para avaliar uma solução produzida para o MMAP, a alocação dos objetos aos grupos é realizada buscando alcançar a máxima homogeneidade interna. Assim, em um conjunto de dados a ser protegido, se houver ao menos c objetos idênticos em termos dos valores dos atributos que os descrevem, esses objetos serão alocados no mesmo grupo e, portanto, nenhuma perturbação será inserida na fase de agregação. Esses objetos não serão protegidos no conjunto de dados resultantes, pois permaneceram representados pelos valores originais. A fim de evitar esta condição indesejável, o risco de violação de sigilo deve também ser controlado.

O risco de violação de sigilo (DR) depende da quantidade de informação adicional que está disponível sobre os objetos representados no conjunto de dados original. Mais especificamente, caso estejam disponíveis informações de T não submetidas a técnicas adequadas de anonimização, métodos de pareamento de registros podem permitir a reidentificação dos objetos (Winkler [2004], Skinner [2008], Domingo-Ferrer e Muralidhar [2016]).

Segundo Domingo-Ferrer e Torra [2001], *Linkage Disclosure* (LD) é o mecanismo padrão usado para medir o DR de um método de SDC. *Distance-based Linkage Disclosure* (DLD), descrito em Herranz et al. [2012] e dado pela Equação 2, é uma medida de LD comumente usada (Yancey et al. [2002], Truta et al. [2004], Mortazavi e Jalili [2015]), pois permite quantificar a dificuldade de identificar um objeto protegido através de técnicas de *record linkage* e *statistical matching*.

$$DLD(T, T') = \frac{\#\{x'_j | j = \arg\min_h(d(x'_j, x_h)), x'_j \in T', x_h \in T\}}{n} \tag{2}$$

Essa medida é calculada considerando o número de atributos originais que o usuário do dado protegido tem acesso, conforme Mortazavi e Jalili [2015]. Com o objetivo de evitar repetir seu cálculo diversas vezes, para diferentes cenários, Herranz et al. [2012] propos assumir o caso mais desfavorável no que diz respeito à preservação do sigilo. Esse caso ocorre quando todos os atributos originais estão disponíveis para o usuário do dado protegido.

No que diz respeito à complexidade computacional, para um conjunto de dados com número de atributos quantitativos superior a dois, a tarefa de definir a partição ótima de seus objetos em grupos homogêneos de no mínimo c objetos pertence à classe dos problemas NP-difíceis, segundo Oganian e Domingo-Ferrer [2001]. Desta forma, não é possível resolver o MMAP através de métodos de enumeração exaustiva e há limitações quanto ao uso de formulações matemáticas.

Assim, no presente trabalho, uma vez definidas as medidas de IL e DR, o MMAP foi abordado como um problema de otimização mono-objetivo e resolvido através do uso de uma heurística híbrida baseada nos conceitos da metaheurística Biased Random Key Genetic Algorithm – BRKGA, proposta por Gonçalves e Resende [2011].

DOI: 10.59254/sbpo-2023-174920
http://live-proceedings.pantheonsite.io/p/174920

Para isso foi elaborada a função objetivo parametrizada descrita pela Equação 3, sujeita às restrições dadas pelas Equações 4, 5, 7 e 6.

$$\text{Minimizar} \sum_{m=1}^{M} \alpha_m F_m(T, T') \tag{3}$$

$$c \leq n_i < 2c, \ \forall \ i \in \{1, ..., k\} \tag{4}$$

$$\sum_{i=1}^{k} n_i = n \tag{5}$$

$$G_r \cap G_s = \emptyset, \ \forall \ r, s \in \{1, ..., k\}, r \neq s \tag{6}$$

$$n_i \in \mathbb{Z}_+, \ \forall \ i \in \{1, ..., k\} \tag{7}$$

Onde $F_m(T, T')$ representa as medidas de perda de informação (IL) e de risco de violação de sigilo (DR) para um par T (conjunto original) e T' (conjunto resultante). Os parâmetros α_m definem a ponderação ou contribuição[2] de $F_m(T, T')$ no valor final da função objetivo, conforme posto em French [2018], sendo $0 \leq \alpha_m \leq 1 \ \forall \ m = 1, ..., M$ e $\sum_{m=1}^{M} \alpha_m = 1$, M correspondente ao número de medidas consideradas na função objetivo parametrizada.

A restrição descrita na Equação 4 define os limites inferior e superior para o número de objetos em cada grupo G_i, em função do nível de agregação c. O limite inferior corresponde ao nível de agregação, c. Quanto ao limite superior, foi demonstrado por Domingo-Ferrer e Mateo-Sanz [2002] que, em uma solução ótima do MMAP, o tamanho dos grupos (n_i) é menor do que $2c$, para todo $i \in \{1, ..., k\}$. As restrições descritas pelas Equações 5 e 6 estabelecem que o particionamento deve ser completo e exclusivo, ou seja, todo objeto do conjunto T deve ser alocado em um e apenas um grupo. A restrição descrita pela Equação 7 é a restrição de integralidade do problema e garante que o número de objetos em qualquer grupo G_i seja um inteiro estritamente positivo.

No que diz respeito à heuristica proposta, foram adotados os operadores de cruzamento e mutação da metaheurística BRKGA (Gonçalves e Resende [2011]). No entanto, uma forma alternativa de representação das soluções foi utilizada. A metaheurística BRKGA utiliza chaves aleatórias para representar os cromossomos associados às soluções de um problema de otimização, obtidas mediante aplicação de um decodificador. Um vetor de chaves aleatórias é um vetor de números reais u gerado segundo uma distribuição uniforme no intervalo $[0, 1]$. Um decodificador é um procedimento determinístico que recebe como entrada um vetor de chaves aleatórias e o transforma em uma solução viável para o problema de otimização em questão.

O decodificador adotado na heurística proposta, descrito no Algoritmo 1, define o tamanho n_i dos grupos iterativamente. Mais especificamente, na iteração i o tamanho n_i do grupo G_i é definido respeitando a restrição $c \leq n_i < 2c$ (Equação 4), e o total de objetos alocados $n_{alocados}$ desde a primeira iteração é acumulado, $n_{alocados} = \sum_{j=1}^{i} n_j$. Esse procedimento se repete até que a diferença entre o total de objetos alocados $n_{alocados}$ até o grupo G_i e o total de objetos n seja menor do que o limite superior de tamanho dos grupos, $2c$. Os objetos restantes são então alocados ao último grupo G_k.

[2] As medidas utilizadas na função objetivo parametrizada devem possuir a mesma escala.

DOI: 10.59254/sbpo-2023-174920
http://live-proceedings.pantheonsite.io/p/174920

Uma vez definidos os tamanhos e número de grupos, é gerado um vetor sequencial de rótulos de grupos, de comprimento n. Esse vetor é então ordenado segundo um vetor de chaves aleatórias, atribuindo rótulos de grupos aos objetos.

Não é possível, no entanto, aplicar o procedimento inverso para recodificar soluções. A ordenação original do vetor de chaves aleatórias não é preservada no processo de decodificação inversa e, uma vez que o tamanho dos grupos é definido pelas k primeiras posições desse vetor, onde k é o número total de grupos da solução resultante, uma solução diferente é obtida.

Algorithm 1: Decodificador

$u \leftarrow$ vetor de chaves aleatórias de comprimento n;
$n_{min} \leftarrow c$; $n_{max} \leftarrow 2c - 1$;
$n_{alocados} \leftarrow 0$;
$i \leftarrow 0$;
while $n_{max} \geq c$ **do**
  $i \leftarrow i + 1$;
  $n[i] \leftarrow round(n_{min} + u[i] * (n_{max} - n_{min}))$;
  $n_{alocados} \leftarrow n_{alocados} + n[i]$;
  $r \leftarrow n - n_{alocados}$;
  $n_{max} \leftarrow \min(n_{max}, r - n_{min})$;
end
if $n_{alocados} < n$ **then**
  $i \leftarrow i + 1$;
  $n[i] \leftarrow n[i] + n - n_{alocados}$;
end
$k_{max} \leftarrow i$;
$j \leftarrow 0$;
for $i = 1$ *to* k_{max} **do**
  while $n[i] > 0$ **do**
    $j \leftarrow j + 1$;
    cromossomo$[j] \leftarrow i$;
    $n[i] \leftarrow n[i] - 1$;
  end
end
ordenar o cromossomo segundo u;
return cromossomo;

Uma vez definido o decodificador, o algoritmo BRKGA é executado, onde, na primeira iteração ($g = 1$) é gerada uma população inicial de p indivíduos, representados por vetores de chaves aleatórias, que são decodificados em cromossomos representando soluções viáveis para o problema, que são avaliadas segundo uma função de *fitness*, a fim de medir a qualidade das soluções associadas a cada cromossomo. Os cromossomos são então divididos em dois conjuntos, um denominado elite, denotado por C_e, reunindo os p_e cromossomos com melhor valor de *fitness*, e outro, denominado não elite, denotado por C_{ne}, reunindo os $p - p_e$ cromossomos restantes, sendo $p_e < p - p_e$.

Para as gerações seguintes, em particular na geração $g+1$, os p_e cromossomos da geração anterior (g) alocados em C_e são copiados para a população da nova geração ($g + 1$) e p_m cromossomos mutantes são gerados através do mesmo procedimento adotado para gerar os cromossomos da população inicial. Os demais $p - p_e - p_m$ cromossomos são descendentes produzidos através da aplicação do operador de cruzamento a pares de cromossomos da população da geração g, de-

DOI: 10.59254/sbpo-2023-174920
http://live-proceedings.pantheonsite.io/p/174920

nominados cromossomos pais, sendo cada par composto por um cromossomo de C_e e um de C_{ne}. A definição do gene (elemento do vetor de chaves aleatórias) que será herdado pelo descendente é feita através de uma seleção viciada, onde ρ_e, sendo $\rho_e > 0.5$, define a probabilidade de cada gene ser herdado do cromossomo pai proveniente de C_e.

4. Experimentos

Esta seção traz os resultados dos experimentos computacionais conduzidos a fim de avaliar o desempenho do método em produzir soluções adequadas para o MMAP, isto é, minimizando simultaneamente o risco de violação de sigilo e a perda de informação, em conjuntos de dados com diferentes características e procedências. Como objetivo secundário, os experimentos buscaram, ainda, demonstrar o comportamento da função objetivo parametrizada no que se refere à redução das medidas de IL e DR, conforme especificado pelo usuário.

O método foi implementado em linguagem R (R Core Team [2021]) aproveitando os recursos da arquitetura *multicore* do computador e as funções de paralelismo disponíveis nos pacotes `parallel` (R Core Team [2021]) e `doParallel` (Corporation e Weston [2022]) do R.

Todos os experimentos foram realizados em computador dotado de um processador i7-8565U de 1.8 GHz e 4 núcleos, 16Gb de RAM, sistema operacional Linux Ubuntu versão 22.04.2 LTS e R versão 4.1.2.

Os resultados produzidos pelo método proposto foram comparados com os resultados de três métodos da literatura, disponíveis através do pacote `sdcMicro` (Templ et al. [2015]), que, em estudo comparativo realizado por Templ et al. [2015], apresentaram bons resultados em termos da redução de IL ou DR. Sejam eles: MDAV (Domingo-Ferrer e Torra [2005]); clustPPPCA[3] (Templ [2006]); e PPPCA (Croux e Ruiz-Gazen [2005]).

A fim de demonstrar a efetividade da função objetivo parametrizada proposta na Equação 3, foram consideradas as medidas de IL e DR dadas pelas Equações 1 e 2, respectivamente $F_1(T, T') = IL(T, T')$ e $F_2(T, T') = DLD(T, T')$, em 3 diferentes cenários, representados pelas seguintes funções objetivo:

$$f_{obj_1} = IL(T, T')$$

$$f_{obj_2} = 0.7 * IL(T, T') + 0.3 * DLD(T, T')$$

$$f_{obj_3} = 0.3 * IL(T, T') + 0.7 * DLD(T, T')$$

Além dos parâmetros referentes à função objetivo, foi necessário ainda configurar cinco parâmetros referentes à execução do BRKGA, quais sejam: g: total de gerações; p: tamanho da população; p_e: tamanho do conjunto elite, em proporção da população; p_m: número de cromossomos mutantes introduzidos em cada geração, em proporção da população; ρ_e: probabilidade de seleção de genes do cromossomo elite no cruzamento. Experimentos preliminares foram realizados a fim de calibrar os valores dos parâmetros, chegando à seguinte configuração: $p_e = 0.1$, $p_m = 0.2$, $\rho_e = 0.9$, $p = 100$ e $g = 100$.

Finalmente, no que diz respeito ao nível de agregação, este parâmetro foi definido como $c = 5$, isto é, durante a contrução das soluções, o número de objetos no i-ésimo grupo, dado por n_i, deve respeitar a restrição $5 \leq n_i < 10,\ \forall\, i \in \{1, ..., k\}$, conforme a Equação 4, uma vez que é demonstrado em Domingo-Ferrer e Mateo-Sanz [2002] que, em uma solução ótima do MMAP, $n_i < 2c\ \forall i \in \{1, ..., k\}$.

[3]No método clustPPCA, o algoritmo de agrupamento CLARA Kaufman e Rousseeuw [1990] foi aplicado antes de realizar a microagregação. O número de grupos na etapa preliminar de agrupamento variou entre os números inteiros contidos no intervalo $[3, 8]$ e a melhor solução obtida, em termos de IL, foi considerada como solução final do método.

DOI: 10.59254/sbpo-2023-174920
http://live-proceedings.pantheonsite.io/p/174920

O método foi aplicado a 16 conjuntos de dados disponíveis publicamente, doravante denominados instâncias, descritas na Tabela 1 em termos dos seus números de atributos, que variam de 2 a 60, e de objetos, que variam de 75 a 1080. As instâncias 1 a 10 são típicas da literatura de análise de agrupamentos, as de 11 a 14 são instâncias sintéticas com distribuição conhecida e as de 15 a 16 são instâncias proveninentes da literatura de SDC[4].

Tabela 1: Características das instâncias utilizadas no experimento computacional.

#	instância	objetos	atributos
1	200DATA	200	2
2	chart	600	60
3	gauss9	900	2
4	iris	150	4
5	maronna	200	2
6	outliers	150	2
7	ruspini	75	2
8	spherical 4d3c	400	3
9	synthetic control	600	51
10	wine	178	13
11	Gamma400	500	3
12	Normal300	300	2
13	Uniform400	400	2
14	Uniform700	700	2
15	Census	1080	13
16	Tarragona	834	13

A fim de dirimir questões relacionadas ao comportamento estocástico dos métodos, tanto para o método proposto, como para os métodos da literatura, foram feitas 5 execuções independentes para cada instância, considerando a mesma configuração de parâmetros. Para efeitos de comparação, foi adotada, para todos os métodos, a solução que apresentou o melhor resultado dentre as 5 execuções.

As Tabelas 2 e 3 apresentam os valores de IL e DL, respectivamente, produzidos pelos métodos considerados.

Tabela 2: Resultados – Perda de informação em termos de $IL(T, T')$.

#	instância	f_{obj_1}	f_{obj_2}	f_{obj_3}	MDAV	clustPPPCA	PPPCA
1	200DATA	0.125	0.150	0.162	**0.006**	0.009	0.019
2	chart	0.713	0.730	0.742	**0.563**	0.646	0.642
3	gauss9	0.496	0.495	0.508	**0.006**	0.047	0.392
4	iris	0.163	0.174	0.145	0.047	**0.034**	0.035
5	maronna	0.224	0.227	0.269	**0.019**	0.032	0.181
6	outliers	0.366	0.364	0.400	0.038	**0.030**	0.041
7	ruspini	0.145	0.120	0.158	**0.005**	0.014	0.019
8	spherical 4d3c	0.207	0.221	0.228	0.220	**0.009**	0.011
9	synthetic control	0.713	0.717	0.741	**0.170**	0.240	0.292
10	wine	0.171	0.190	0.233	**0.049**	0.136	0.419
11	Gamma400	0.533	0.523	0.547	**0.037**	0.107	0.440
12	Normal300	**0.447**	0.489	0.510	0.537	0.637	0.639
13	Uniform400	0.524	0.497	0.540	**0.012**	0.093	0.371
14	Uniform700	0.594	0.562	0.586	**0.006**	0.066	0.374
15	Census	0.558	0.568	0.572	0.137	**0.078**	0.178
16	Tarragona	0.390	0.355	0.377	0.195	0.171	**0.141**

No que diz respeito a IL (Tabela 2), o MDAV apresentou o melhor desempenho, encontrando a solução de menor IL para 10 das 16 instâncias, seguido pelo clustPPPCA, com 4 instâncias, e pelo PPPCA, com apenas 1 instância. O método proposto, utilizando a f_{obj_1}, também alcançou o melhor resultado para apenas 1 instância.

[4]O menor número de instâncias provenientes da literatura de SDC se deve ao fato de que, em geral, os estudos de caso de SDC envolvem dados não públicos. Com isso, a disponibilidade de conjuntos de dados é relativamente restrita, ainda mais quando considerados apenas atributos quantitativos.

DOI: 10.59254/sbpo-2023-174920
http://live-proceedings.pantheonsite.io/p/174920

Tabela 3: Resultados – Risco de violação de sigilo em termos de $DLD(T, T')$.

#	instância	f_{obj_1}	f_{obj_2}	f_{obj_3}	MDAV	clustPPPCA	PPPCA
1	200DATA	0.015	**0**	**0**	0.195	0.160	0.150
2	chart	0.043	0.003	**0.002**	0.110	0.055	0.053
3	gauss9	0.004	**0**	**0**	0.194	0.113	0.052
4	iris	0.073	**0**	**0**	0.173	0.173	0.127
5	maronna	0.045	0.005	**0**	0.200	0.160	0.115
6	outliers	0.038	0.015	**0**	0.173	0.200	0.187
7	ruspini	0.107	**0**	**0**	0.182	0.100	0.092
8	spherical 4d3c	0.013	0.002	**0**	0.028	0.185	0.197
9	synthetic control	0.037	0.002	**0**	0.198	0.160	0.160
10	wine	0.028	0.006	**0**	0.186	0.108	0.030
11	Gamma400	0.006	**0**	0.002	0.190	0.167	0.083
12	Normal300	0.013	0.007	**0**	0.123	0.053	0.048
13	Uniform400	0.010	0.005	**0**	0.190	0.140	0.050
14	Uniform700	0.003	0.003	**0**	0.196	0.114	0.056
15	Census	0.004	0.004	**0**	0.095	0.091	0.043
16	Tarragona	0.008	0.005	**0.001**	0.101	0.054	0.064

O método proposto, ao utilizar as f_{obj_2} e f_{obj_3}, não produziu nenhuma melhor solução em termos de IL. Este resultado era esperado, uma vez que ambas as funções objetivo consideram não apenas IL, mas também DLD na busca das soluções. Por se tratarem de objetivos concorrentes, isto é, a redução de IL implica no aumento de DLD e vice-versa, é improvável que soluções produzidas a partir dessa estratégia produza soluções competitivas em relação aos métodos da literatura considerados.

No que diz respeito ao desempenho inferior da f_{obj_1}, que considera apenas IL, a principal hipótese é que, por contar com inicialização aleatória, isto é, inicialmente os objetos são alocados aos grupos desconsiderando os valores dos seus atributos, o método proposto favorece a construção de soluções com menores valores de DLD e, consequentemente, maiores valores de IL. Não foram realizados experimentos complementares para testar tal hipótese, mas os autores acreditam que o aumento do número de gerações, assim como um processo de calibração de parâmetros específico para a execução do método adotanto f_{obj_1}, pode oferecer resultados mais competitivos em termos de IL, frente aos outros métodos considerados.

Os resultados em termos de DLD, apresentados na Tabela 3, corroboram com tal hipótese, pois permitem verificar que f_{obj_1} alcançou patamares de DLD em geral bem inferiores aos das soluções produzidas pelos métodos da literatura. Entretanto, como esperado, f_{obj_2} e f_{obj_3} dominaram as melhores soluções em termos desta medida. Em particular, ao utilizar f_{obj_3}, o método proposto produziu a melhor sulução, em termos de DLD, para 15 das 16. Todavia, é necessário considerar que, para 13 delas, o valor foi igual a zero. Este resultado, ainda que interessante em termos da garantia do sigilo, sugere que pode haver espaço para melhorar as soluções produzidas em termos de IL. Neste sentido, o uso da f_{obj_2} produziu soluções mais equilibradas.

Cabe ainda destacar que os resultados demonstraram a efetividade da função objetivo parametrizada, no que diz respeito à ponderação das medidas de interesse. Em termos de IL, as soluções produzidas quando adotada f_{obj_1} superam, em geral, aquelas produzidas utilizando a f_{obj_2}, que, por sua vez, superam aquelas produzidas através da f_{obj_3}. No que diz respeito a DLD, observa-se o comportamento inverso, como esperado. Naturalmente, a calibração e especificação dos parâmetros e medidas utilizadas na função objetivo é de crucial importância para obtenção do resultado desejado.

DOI: 10.59254/sbpo-2023-174920
http://live-proceedings.pantheonsite.io/p/174920

5. Conclusões

O método proposto superou os métodos da literatura considerados, no que diz respeito ao controle do risco de violação de sigilo, medido em termos de DLD. Por este objetivo ser conflitante com a redução da perda de informação, os resultados produzidos em termos de IL foram, geralmente, inferiores aos dos demais métodos. Entretando, em função do método proposto adotar função objetivo especificada pelo usuário, tal comportamento pode ser ajustado para produzir soluções especialmente adequadas a aplicações específicas. A efetividade deste recurso, garantida pela flexibilidade da função objetivo, foi demonstrada nos resultados produzidos na realização do experimento. Ainda neste sentido, a configuração dos parâmetros do BRKGA também pode ser explorada pelo usuário a fim de contribuir no alcance deste objetivo.

No que diz respeito aos tempos de execução, embora não apresentados numericamente, o desempenho do método proposto foi inferior ao dos demais métodos neste aspecto, em função principalmente de ser baseado em uma metaheurística que utiliza procedimentos de construção e busca local que, por sua vez, demandam computação intensiva. Além disso, é custoso o cálculo da medida de DLD, adotada nas funções objetivo f_{obj_2} e f_{obj_3} e, portanto, realizado para todas as soluções (tamanho da população) a cada geração. Ainda neste sentido, recursos de paralelismo foram empregados para mitigar este último aspecto, fazendo com que os tempos de execução variassem entre 1.6 e 25.8 segundos, a depender do número de objetos e atributos do conjunto de dados. Nos demais métodos, o tempo máximo de execução observado foi de inferior a 1 segundo, sendo MDAV o mais eficiente.

O método proposto se mostrou efetivo em resolver o MMAP de maneira satisfatória e oferece flexibilidade para produzir soluções potencialmente mais adequadas a diferentes aplicações. Em relação ao tempo de execução, embora tenha apresentado valores superiores aos dos demais métodos, estes não devem representar impedimento ao seu uso na grande maioria das aplicações.

Referências

Corporation, M. e Weston, S. *doParallel: Foreach Parallel Adaptor for the 'parallel' Package*, 2022. URL `https://CRAN.R-project.org/package=doParallel`. R package version 1.0.17.

Croux, C. e Ruiz-Gazen, A. (2005). High breakdown estimators for principal components: the projection-pursuit approach revisited. *Journal of Multivariate Analysis*, 95(1):206–226.

Domingo-Ferrer, J. e Mateo-Sanz, J. M. (2002). Practical data-oriented microaggregation for statistical disclosure control. *IEEE Trans. on Knowl. and Data Eng.*, 14(1):189–201. ISSN 1041-4347.

Domingo-Ferrer, J., Martínez-Ballesté, A., Mateo-Sanz, J. M., e Sebé, F. (2006). Efficient multivariate data-oriented microaggregation. *The VLDB Journal*, 15(4):355–369. ISSN 1066-8888.

Domingo-Ferrer, J. e Muralidhar, K. (2015). New directions in anonymization: Permutation paradigm, verifiability by subjects and intruders, transparency to users. *CoRR*.

Domingo-Ferrer, J. e Muralidhar, K. (2016). New directions in anonymization: Permutation paradigm, verifiability by subjects and intruders, transparency to users. *Information Sciences*, 337-338:11 – 24.

Domingo-Ferrer, J. e Torra, V. (2001). *A quantitative comparison of disclosure control methods for microdata*, p. 111–133. Elsevier.

Domingo-Ferrer, J. e Torra, V. (2005). Ordinal, continuous and heterogeneous k-anonymity through microaggregation. *Data Min. Knowl. Discov.*, 11(2):195–212.

Elliot, M. e Domingo-Ferrer, J. (2018). The future of statistical disclosure control. *CoRR*, abs/1812.09204. URL `http://arxiv.org/abs/1812.09204`.

DOI: 10.59254/sbpo-2023-174920
http://live-proceedings.pantheonsite.io/p/174920

Fayyoumi, E. e Oommen, B. J. (2010). A survey on statistical disclosure control and microaggregation techniques for secure statistical databases. *Softw. Pract. Exper.*, 40(12):1161–1188. ISSN 0038-0644.

French, M. (2018). *Fundamentals of Optimization: Methods, Minimum Principles, and Applications for Making Things Better*. Springer International Publishing.

Gonçalves, J. e Resende, M. G. C. (2011). Biased random-key genetic algorithms for combinatorial optimization. *Journal of Heuristics*, 17(5):487–525.

Herranz, J., Nin, J., e Solé, M. (2012). Kd-trees and the real disclosure risks of large statistical databases. *Inf. Fusion*, 13(4):260–273. ISSN 1566-2535.

Jolliffe, I. (2002). *Principal Component Analysis*. Springer Series in Statistics. Springer.

Kaufman, L. e Rousseeuw, P. (1990). *Finding Groups in Data: an introduction to cluster analysis*. Wiley.

Mortazavi, R. e Jalili, S. (2015). Preference-based anonymization of numerical datasets by multi-objective microaggregation. *Inf. Fusion*, 25(C):85–104.

Oganian, A. e Domingo-Ferrer, J. (2001). On the complexity of optimal microaggregation for statistical disclosure control. *Statistical Journal of the United Nations Economic Comission for Europe*, 18:345–354.

ONU. *United Nations Guide on Privacy-Enhancing Technologies for Official Statistics*. United Nations Committee of Experts on Big Data and Data Science for Official Statistics, New York, 2023. Disponível em: `https://unstats.un.org/bigdata/task-teams/privacy/guide/index.cshtml`.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021. URL `https://www.R-project.org/`.

Skinner, C. (2008). Assessing disclosure risk for record linkage. In Domingo-Ferrer J., T. V., editor, *Privacy in Statistical Databases. PSD 2008. Lecture Notes in Computer Science*, volume 5262, p. 166–176. Springer, Berlin, Heidelberg.

Solanas, A., Martínez-Ballesté, A., Mateo-Sanz, J., e Domingo-Ferrer, J. (2006). Multivariate microaggregation based genetic algorithms. *3rd International IEEE Conference Intelligent Systems*, p. 65–70.

Solanas, A. (2008). *Privacy Protection with Genetic Algorithms*, p. 215–237. Springer Berlin Heidelberg, Berlin, Heidelberg.

Templ, M. (2006). Software development for sdc in r. In Domingo-Ferrer, J. e Franconi, L., editors, *Privacy in Statistical Databases*, p. 347–359. Springer Berlin Heidelberg. ISBN 978-3-540-49332-7.

Templ, M., Kowarik, A., e Meindl, B. (2015). Statistical disclosure control for micro-data using the R package sdcMicro. *Journal of Statistical Software*, 67(4):1–36.

Truta, T. M., Fotouhi, F., e Barth-Jones, D. (2004). Assessing global disclosure risk in masked microdata. In *Proceedings of the 2004 ACM Workshop on Privacy in the Electronic Society*, WPES '04, p. 85–93, New York, NY, USA. ACM. ISBN 1-58113-968-3.

Winkler, W. E. (2004). Re-identification methods for masked microdata. In Domingo-Ferrer J., T. V., editor, *Privacy in Statistical Databases. PSD 2004. Lecture Notes in Computer Science*, volume 3050, p. 231–246. Springer, Berlin, Heidelberg.

Yancey, W. E., Winkler, W. E., e Creecy, R. H. (2002). Disclosure risk assessment in perturbative microdata protection. In *Inference Control in Statistical Databases, From Theory to Practice*, p. 135–152, London, UK, UK. Springer-Verlag.

Agradecimentos: Ao CNPq - Projeto Universal 405044/2021-6.

DOI: 10.59254/sbpo-2023-174920
http://live-proceedings.pantheonsite.io/p/174920