PROPOSTA DE MÉTODO HEURÍSTICO APLICADO AO PROBLEMA DE MICROAGREGAÇÃO MULTIVARIADO

Augusto César Fadel
Instituto Brasileiro de Geografia e Estatística - IBGE
Avenida República do Chile, 500, Centro, Rio de Janeiro, RJ
augustofadel@gmail.com

Luiz Satoru Ochi
Universidade Federal Fluminense - IC/UFF
Avenida General Milton Tavares de Souza, s/n, São Domingos, Niterói, RJ
luiz.satoru@gmail.com

José André de Moura Brito
Escola Nacional de Ciências Estatísticas - ENCE/IBGE
Rua André Cavalcante, 106, Bairro de Fátima, Rio de Janeiro, RJ
jambrito@gmail.com

Gustavo Silva Semaan
Instituto do Noroeste Fluminense de Educação Superior - INFES/UFF
Avenida João Jasbick, s/n, Bairro Aeroporto, Santo Antônio de Pádua, RJ
gustavosemaan@gmail.com

RESUMO

A divulgação de microdados por parte de instituições públicas, em especial Institutos Nacionais de Estatística, e o compartilhamento entre organizações públicas e privadas e a academia são de importância inquestionável nos dias atuais. Entretanto, aspectos éticos e legais, referentes à preservação do sigilo do indivíduo, seja ele pessoa física ou jurídica, muitas vezes impossibilitam esta cooperação. Neste sentido, os métodos de Controle Estatístico de Sigilo, mais especificamente, o método de microagregação, oferecem alternativa para viabilizar a divulgação e o compartilhamento de microdados, através do controle do risco de violação do sigilo. O presente artigo propõe novo método heurístico, utilizando conceitos da metaheurística BRKGA, e traz um comparativo desse com métodos considerados de bom desempenho, no que diz respeito à redução da perda de informação. Os resultados computacionais em bases reais e da literatura indicam que o método é competitivo, superando métodos clássicos no que diz respeito à redução da perda de informação e do risco de violação de sigilo.

PALAVRAS CHAVE. Controle de Sigilo, Microagregação, Metaheurísticas.

Tópicos EST, MH, OC

ABSTRACT

The dissemination of microdata by public institutions, especially National Statistical Offices, and the sharing of data among public and private organizations and the academy are of undeniable importance today. However, ethical and legal aspects, regarding privacy protection, often make such cooperation impossible. Methods of Statistical Disclosure Control, more specifically the microaggregation technique, offer an alternative to enable the dissemination and sharing of microdata. This paper proposes a new heuristic method based on the concepts of BRKGA metheuristic and presents a comparative study with other methods. The computational results show that the method is competitive, overcoming classic methods both in terms of reducing information loss and disclosure risk.

KEYWORDS. Disclosure Control. Microaggregation. Metaheuristics.

Paper topics EST, MH, OC

1. Introdução

O compartilhamento de dados e, consequentemente, o controle de sigilo, são temas de grande importância atualmente. 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 [UNSTATS 2017b]. Entretanto, aspectos éticos e legais, relacionados à garantia do sigilo do informante, restringem a divulgação ou disponibilização de tais informações. Organizações privadas, tais como empresas de telefonia móvel, comércio varejista e companhias aéreas, por sua vez, detêm bases de dados úteis para a elaboração de políticas públicas [UNSTATS 2017a], que não podem ser compartilhadas principalmente em função de interesses comerciais. Neste sentido, as técnicas modernas de controle de sigilo podem viabilizar a divulgação de dados por parte dos INE e o compartilhamento de dados entre organizações públicas e privadas.

As técnicas de controle de sigilo hoje disponíveis podem ser divididas em três categorias, sejam elas: i) servidores de modelagem e preservação de confidencialidade (do inglês, *confidentiality preserving model servers*); ii) Instalações de acesso remoto ou salas de acesso restrito; e iii) controle estatístico de sigilo (do inglês, *Statistical Disclosure Control* - SDC).

Os servidores de modelagem são eficazes na preservação da confidencialidade, mas não oferecem flexibilidade, pois poucos métodos estão disponíveis nesses servidores. Além disso, sem acesso aos dados originais, é problemático avaliar o desempenho (i.e. ajuste) de um modelo. Esta alternativa é extremamente restritiva e incompatível com métodos modernos de produção de estatísticas [Templ 2006]. As salas de acesso restrito são mais flexíveis, no sentido de que o pesquisador pode explorar o dado e especificar um método para ser aplicado. A confidencialidade é garantida por um processo de auditoria realizado pelo proprietário do dado (e.g. INE), antes que o pesquisador tenha acesso ao resultado final. Ambas as alternativas são limitadas e custosas em termos de tempo e recursos financeiros.

Considerando estas questões, uma abordagem mais moderna pode ser realizada através de métodos de SDC. O principal objetivo é modificar o conjunto de dados original reduzindo a quantidade de informação nele disponível, a fim de garantir o sigilo. Estes métodos podem ser divididos em: i) métodos não perturbativos, nos quais apenas o nível de detalhe da informação é alterado através, por exemplo, da agregação de categorias (generalização e supressão); ii) métodos perturbativos, nos quais o dado original é modificado, inserindo algum tipo de erro; e iii) geração sintética de dados, nos quais os dados originais são substituídos por dados sintéticos gerados segundo um modelo. A qualidade do conjunto resultante deve ser avaliada através de duas dimensões concorrentes, sejam elas: a perda de informação e o risco de violação de sigilo. A primeira dimensão faz referência à utilidade do conjunto produzido e a segunda está relacionada com a garantia do sigilo. Um comparativo envolvendo diversos métodos de SDC [Domingo-Ferrer et al. 2001] apontou a microagregação multivariada, método perturbativo, como de maior eficácia quanto à perda de informação e à garantia do sigilo.

A microagregação consiste em particionar os n registros (ou objetos) de uma base de dados em pequenos grupos homogêneos de tamanho n_i, estando n_i compreendido em $c \leq n_i < 2c \; \forall \; i \in \{1, ..., \lfloor n/c \rfloor\}$, sendo c um inteiro positivo correspondente ao nível de agregação. O critério de homogeneidade é dado por uma métrica (e.g. distância euclidiana), que permite avaliar a similaridade entre os objetos da base. Os atributos dos objetos são então substituídos pelo agregado correspondente (e.g. média do grupo) e esses valores são divulgados em substituição aos valores originais [Elliot et al. 2005].

Além da aplicação de controle estatístico de sigilo, para a qual foi concebida, a técnica de microagregação encontra aplicações em outras áreas, como inteligência artificial [Domingo-Ferrer e Torra 2003] e mineração de dados [Solanas et al. 2006], por exemplo.

No aspecto computacional, o problema de microagregação univariado, ou seja, quando apenas um atributo é considerado, tem similaridade com o problema do caminho mínimo e pode ser resolvido em tempo polinomial [Hansen e Mukherjee, 2003]. Entretanto, o problema de microagregação multivariado é NP-difícil [Oganian e Domingo-Ferrer 2001], o que impossibilita a aplicação de um método de enumeração exaustiva para resolvê-lo. A importância do tema

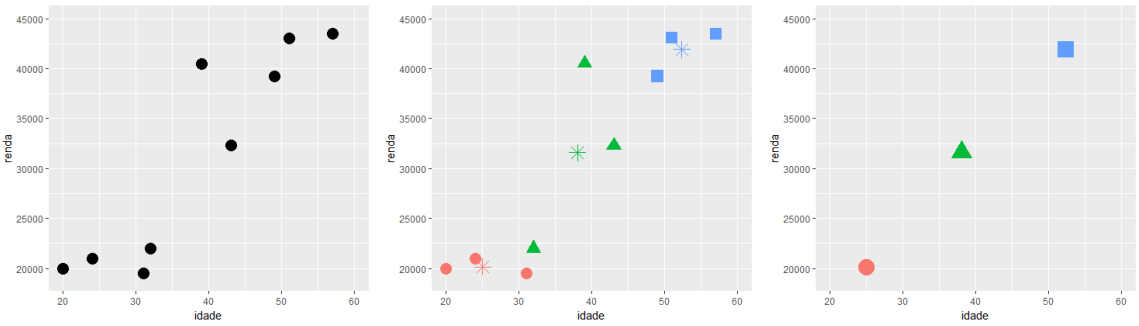
conjugada à inviabilidade de abordagens exatas têm motivado, nos últimos tempos, o desenvolvimento de uma grande variedade de métodos, aplicando heurísticas e metaheurísticas. O presente artigo apresenta comparativo envolvendo métodos da literatura e propõe novo método, que utiliza conceitos da metaheurística algoritmo genético de chaves aleatórias viciadas (do inglês, *Biased Random Key Genetic Algorithm* - BRKGA) [Gonçalves e Resende 2011].

Esse artigo conta com cinco seções, além da introdução. Na Seção 2 é apresentada uma descrição formal do problema e a Seção 3 traz uma breve revisão da literatura. O método proposto é descrito na Seção 4. A Seção 5 apresenta os resultados do experimento computacional realizado, no qual o método proposto e outros quatro métodos foram aplicados em 34 instâncias de teste. As soluções produzidas foram avaliadas quanto à perda de informação e ao risco de violação de sigilo e, segundo ambos os critérios, o método proposto superou os demais métodos. A Seção 6 detalha as conclusões obtidas e apresenta propostas para trabalhos futuros.

2. Descrição do problema

O problema de microagregação clássico deve ser resolvido em duas etapas [Domingo-Ferrer et al. 2006], sejam elas: (i) Particionamento: dado um nível de agregação c, a base de dados composta por n registros (ou objetos) é particionada em grupos de tal maneira que objetos alocados em um mesmo grupo sejam similares entre si, segundo uma métrica definida, e que cada grupo tenha, ao menos, c objetos; (ii) Agregação: um operador de agregação (e.g. média) é utilizado para computar o centroide de cada grupo. Cada objeto é então substituído pelo centroide do grupo em que está alocado.

A Figura 1 ilustra o processo de microagregação em uma base de dados hipotética, com 9 objetos e 2 atributos, apresentada em (a). Para o nível de agregação c = 3, os 9 objetos foram particionados em 3 grupos de 3 objetos cada, conforme representado em (b), onde os objetos alocados aos grupos 1, 2 e 3 são indicados, respectivamente, por círculos vermelhos, triângulos verdes e quadrados azuis. Ainda em (b), os asteriscos correspondem ao centróide de cada grupo, na cor correspondente ao grupo. Finalmente, em (c), cada objeto tem o valor de seus atributos substituído pelos valores correspondentes ao centróide de grupo.



(a) base original (b) particionamento (c) agregação

Figura 1 - Exemplo de aplicação do processo de microagregação em base de dados hipotética de 9 objetos, para c = 3

O problema pode ser formalizado como: dado um conjunto de dados T, com f atributos numéricos contínuos e n objetos, k grupos de tamanho mínimo c devem ser construídos de forma a minimizar a perda de informação. A perda de informação ocorre na fase de agregação e é acentuada quando, na fase de particionamento, objetos pouco similares são alocados em um mesmo grupo. Segundo Fayyoumi e Oommen (2010), a solução ótima para o problema de microagregação é aquela na qual a máxima homogeneidade interna dos grupos é obtida. Dessa forma, é comumente utilizado como função objetivo o erro quadrático total (do inglês, *sum of squared errors* - SSE), que calcula o somatório dos quadrados das diferenças dos objetos e seus respectivos centróides, dado pela Equação (1), sendo x_{ij} correspondente ao objeto j do grupo G_i, n_i a quantidade de objetos do grupo G_i e $\bar{x}_i$ é o centroide do grupo G_i, $i \in \{1, ..., k\}$ e $j \in \{1, ..., n_i\}$.

$$SSE = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)' (x_{ij} - \bar{x}_i) \quad (1)$$

A fim de obter uma solução adequada, as seguintes restrições devem ser consideradas:

$$c \leq n_i < 2c \; \forall \; i \in \{1, \ldots, k\} \quad (2)$$

$$\sum_{i=1}^{c} n_i = n \quad (3)$$

$$G_r \cap G_s = \emptyset \; \forall \; r, s \in \{1, \ldots, k\}, r \neq s \quad (4)$$

A restrição (2) limita o número de objetos em cada grupo. Segundo Domingo-Ferrer e Mateo-Sanz (2002), na solução ótima do problema de microagregação o tamanho dos grupos (n_i) é menor do que 2c, para todo i ∈ {1, ..., k}, sendo c o nível de agregação. As restrições (3) e (4) indicam que todo objeto deve ser alocado em um e somente um grupo.

3. Revisão da literatura

O problema de microagregação pode ser mapeado em problemas conhecidos de otimização combinatória. Além do problema de agrupamento, com o qual há correspondência direta, foram propostas abordagens baseadas nos problemas de encontrar uma árvore geradora mínima [Laszlo e Mukherjee 2005] e do Caixeiro Viajante [Heaton 2011].

A abordagem mais simples para o problema de microagregação consiste em ordenar os objetos segundo uma variável e, em seguida, particioná-los em grupos de tamanho fixo c. Se o número de objetos não for múltiplo de c, o último grupo assume tamanho superior a c e inferior a 2c. Tal abordagem tem baixo custo computacional, mas os resultados obtidos não são de boa qualidade quanto à preservação da informação e à garantia do sigilo. Podem, no entanto, ser utilizados como linha de base na comparação entre métodos. *Single* [Templ 2006] é um exemplo de método que adota essa abordagem. A fim de obter soluções baseadas em ordenação dos objetos considerando mais de uma variável, é possível aplicar técnicas de redução de dimensionalidade, tal como a técnica de análise de componentes principais (do inglês, *Principal Components Analysis* - PCA) [Jolliffe, 2002]. A ordenação dos objetos é então feita segundo a primeira componente principal [Schmid 2009].

Em Solanas et al. (2006), um método baseado no algoritmo genético clássico [Mitchell 1998] foi proposto e apresentou desempenho superior a métodos clássicos como o MDAV (do inglês, *Maximum Distance to Average Vector*) [Domingo-Ferrer e Torra 2005] (descrito a seguir). Porém, apenas para conjuntos de dados de até 50 objetos. Solanas (2008) apresenta uma abordagem híbrida, conjugando MDAV e algoritmo genético.

A seguir são relacionados métodos de microagregação adotados no comparativo apresentado na Seção 5. Uma revisão mais abrangente pode ser encontrada em Domingo-Ferrer et al. (2006).

O MDAV [Domingo-Ferrer e Torra 2005] é o método de microagregação mais utilizado, segundo Solanas (2008). Esse método consiste em, dado um conjunto de dados, calcular o centróide global (considerando todos os objetos e os seus respectivos atributos) e, em seguida, identificar o objeto (r) de menor similaridade com o centróide global. Dado o nível de agregação c, o objeto r, juntamente com os c - 1 objetos mais similares a ele (em relação aos atributos), dá origem ao primeiro grupo. O segundo grupo é construído de forma análoga, porém tendo como origem o objeto de menor similaridade com o objeto r. O centróide global é então atualizado considerando apenas os objetos ainda não alocados a nenhum grupo e o processo se repete até que todos os objetos tenham sido alocados a algum grupo. A medida de similaridade normalmente adotada é a distância euclidiana.

O método RMDM (do inglês, *Robust Mahalanobis Distance based Microagregation*) [Templ 2007] é uma adaptação do método MDAV. O procedimento executado é o mesmo, porém os centróides são obtidos através de uma medida robusta (e.g. mediana L1) e a similaridade entre os objetos é dada pela distância de Mahalanobis.

Como mencionado, é possível aplicar PCA ao problema de microagregação, porém dois aspectos devem ser considerados. O primeiro se refere ao fato de tal técnica ser sensível à presença de valores extremos (*outliers*), o que pode ser contornado com o uso de métodos robustos para estimar a matriz de covariância [Maronna e Zamar 2002]. O segundo aspecto se refere ao fato de que apenas a primeira componente principal é utilizada no contexto da microagregação, embora

todas sejam estimadas, implicando cálculos redundantes. O método PPPCA (do inglês, *Projection Persuit Principal Components Analysis*) [Croux e Ruiz-Gazen 2005] contorna ambas as desvantagens mencionadas estimando, através de método robusto, apenas a primeira componente, utilizando a técnica de *projection persuit* [Li e Chen 1985].

Para os métodos baseados em ordenação, como o PPPCA, alguns autores propõem realizar uma etapa preliminar na qual um algoritmo de agrupamento é executado e, em seguida, o método de ordenação é aplicado de maneira independente em cada um dos grupos formados. O particionamento dos objetos é então realizado segundo a ordenação estabelecida em cada grupo. Exemplos de métodos que usam esse recurso são o clustPPPCA [Templ 2006], que, após realizar o agrupamento, aplica o método PPPCA em cada grupo, e o método *influence* [Templ 2006], que ordena os objetos de cada grupo obtido na etapa de agrupamento segundo a variável de maior influência, identificada através de método de seleção de variáveis [Liu e Motoda 1998].

Os métodos apresentados foram selecionados com base em estudo realizado por Templ (2018). Os métodos MDAV, RMDM e clustPPPCA produziram soluções de boa qualidade tanto no que diz respeito à perda de informação como em relação ao risco de violação do sigilo. O RMDM, no entanto, não é recomendado para grandes bases de dados, devido ao elevado custo computacional para produzir a matriz de covariância robusta.

4. Método proposto

O presente artigo propõe um novo método de microagregação multivariado utilizando conceitos da metaheurística algoritmo genético de chaves aleatórias viciadas (do inglês, *Biased Random Key Genetic Algorithm* – BRKGA) [Gonçalves e Resende 2011]. São descritos a seguir: i) o esquema de representação das soluções; ii) o processo de inicialização da população de cromossomos (soluções iniciais); iii) a função objetivo para cálculo do ajuste (*fitness*) dos cromossomos; iv) a estratégia de seleção dos cromossomos mais aptos; v) os operadores genéticos de mutação e cruzamento (*crossover*).

O BRKGA parte, em sua geração inicial, de uma população de p cromossomos, representados por vetores cujas componentes têm valores reais gerados aleatoriamente, segundo uma distribuição uniforme [0,1]. É necessário definir um procedimento decodificador para transformar esses vetores em soluções viáveis associadas ao problema em questão. No problema de microagregação, uma solução para a etapa de particionamento deve conter a informação de quais objetos pertencem ao mesmo grupo. Assim sendo, é natural atribuir a cada objeto um rótulo indicando a qual grupo ele pertence. Esse esquema de representação das soluções foi utilizado para propor um método híbrido. Pela definição do problema, cada grupo deve conter, ao menos, c objetos, sendo o número máximo de grupos dado por $k_{max} = \lfloor n/c \rfloor$. Cada solução foi então representada por um cromossomo com n genes, através de um alfabeto de cardinalidade k_{max} no qual cada símbolo (inteiros contidos no intervalo $[1, k_{max}]$) representa um grupo. Esta estrutura pode ser implementada através de um vetor de comprimento n, onde a posição i armazena o grupo ao qual o i-ésimo objeto pertence, $1 \leq i \leq n$.

De forma a atender a restrição (2), para gerar a população inicial foi adotada uma heurística construtiva baseada em soluções do problema do Caixeiro Viajante Euclidiano (PCVE). Cada objeto do conjunto de dados correspondeu a um ponto no espaço f-dimensional, as similaridades foram determinadas através da distância euclidiana, dados os f atributos, e o método *farthest insert* [Rosenkrantz et al., 1977] foi aplicado. Uma vez que o PCV implica retorno à origem, a solução para esse problema é representada por um ciclo, que foi convertido em um caminho através da eliminação da aresta de maior comprimento (peso). Cada um dos p cromossomos da população inicial foi então gerado a partir da divisão desse caminho em segmentos com n_i objetos, onde $c \leq n_i < 2c \; \forall \; i \in [1, k_{max}]$. Os objetos foram rotulados com números inteiros, de maneira que aqueles que compartilhassem o mesmo segmento recebessem o mesmo rótulo, produzindo, portanto, apenas candidatos viáveis.

Quanto ao ajuste (*fitness*) dos cromossomos, medidas associadas à perda de informação podem ser utilizadas. A rigor, medidas de violação de sigilo também podem ser consideradas. No

entanto, o uso de tais medidas tende a produzir soluções com perda de informação elevada. O ajuste foi avaliado através da medida de perda de informação IL1, descrita na Equação (5).

$$\mathrm{IL1} = \frac{\mathrm{SSE}}{\mathrm{SST}} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)'(x_{ij}-\bar{x}_i)}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{x})'(x_{ij}-\bar{x})} \times 100 \qquad (5)$$

Essa medida representa a perda de informação normalizada e está relacionada ao grau de homogeneidade dos grupos obtidos no processo de microagregação. Uma vez obtido o valor de ajuste de cada cromossomo, a população (p) foi dividida em dois conjuntos. O primeiro, chamado conjunto elite, reúne os p_e cromossomos mais aptos, ou seja, com menor valor de IL1. Os demais $p - p_e$ cromossomos compõem o conjunto não elite.

Para cada geração, em particular na geração g + 1, o conjunto elite da geração g é copiado integralmente e p_m novos cromossomos (mutantes) são gerados através do procedimento usado na inicialização da população, a fim de manter a diversidade. Na geração g + 1 a população é então composta por p_e cromossomos do conjunto elite, p_m cromossomos mutantes, e $p - p_e - p_m$ cromossomos, denominados filhos, obtidos a partir do processo de cruzamento.

Mais especificamente, cada filho é gerado a partir do cruzamento entre um cromossomo do conjunto elite e um cromossomo do conjunto não elite, conferindo maior probabilidade de seleção a cromossomos do conjunto elite[1]. O cruzamento é operacionalizado através da construção de um vetor auxiliar v_a de tamanho n, preenchido com valores reais gerados segundo uma distribuição uniforme no intervalo [0,1]. São então pareados o vetor v_a e os dois cromossomos pais selecionados para o cruzamento, quais sejam: I_e pertencente ao conjunto elite e I_n pertencente ao conjunto não elite. Quando a i-ésima posição de v_a for menor ou igual ao parâmetro ρ_e, $\rho_e > 0.5$, o cromossomo filho, I_f, herda, na posição i, o i-ésimo gene de I_e. Caso contrário, herda o i-ésimo gene de I_n. O parâmetro ρ_e define, portanto, a probabilidade de um gene do cromossomo pai proveniente do conjunto elite ser herdado pelo cromossomo filho. Assumindo exemplo hipotético de um conjunto de dados com 10 objetos e $\rho_e = 0.7$, o processo de cruzamento é representado a seguir:

I_e = { **2** , 2 , **1** , **3** , **3** , 1 , **1** , 3 , 1 , **2** }
I_n = { 1 , **1** , 3 , 2 , 2 , **2** , 3 , **3** , **3** , 2 }
v_a = {0.58, 0.89, 0.22, 0.43, 0.70, 0.99, 0.43, 0.75, 0.88, 0.58}
I_f = { 2 , 1 , 1 , 3 , 3 , 2 , 1 , 3 , 3 , 2 }

Esse processo não garante a produção de indivíduos filhos correspondentes a soluções viáveis, pois pode haver grupos com número de objetos maior ou igual a 2c ou menor do que c. Nesse caso, é executado um processo de regeneração, no qual os grupos menores recebem objetos dos grupos maiores, até que a restrição (2) seja satisfeita. O processo de cruzamento com regeneração é repetido até que $p - p_e - p_m$ cromossomos filhos viáveis tenham sido gerados e a população da geração seguinte, g+1, esteja completa.

5. Resultados computacionais

Foi realizado experimento de forma a comparar o desempenho dos métodos MDAV, RMDM, PPPCA e clustPPPCA, bem como do método proposto, doravante referenciado como MABRKGA. Para isso, foram utilizadas 34 instâncias de teste, descritas na Tabela 1. O experimento foi realizado em computador dotado de um processador i7 de 2.6 GHz e 4 núcleos, 8GB de RAM e sistema operacional MAC OS, versão 10.13.3.

As instâncias foram organizadas em quatro subconjuntos, quais sejam: DS1, com instâncias típicas da literatura de análise de agrupamentos; DS2, instâncias sintéticas com distribuição conhecida; DS3, instâncias reais obtidas a partir dos dados do Censo Demográfico 2010, do Produto Interno Bruto (PIB) municipal 2010 e do Índice de Desenvolvimento Humano (IDH) municipal 2010; e DS4, com instâncias típicas da literatura de SDC.

[1] Dado que há p_e indivíduos elite e $p - p_e$ indivíduos não elite, onde $p_e < p - p_e$, a probabilidade de um indivíduo do conjunto elite participar de um cruzamento, dada por $1/p_e$, é superior à probabilidade de um indivíduo do conjunto não elite ser incluído, dada por $1/(p - p_e)$.

Tabela 1 - Instâncias

#	instâncias	objetos	atributos	#	instâncias	objetos	atributos
1	DS1-200DATA	200	2	18	DS3-censo_PA	128	5
2	DS1-chart	600	60	19	DS3-censo_RJ	88	5
3	DS1-gauss9	900	2	20	DS3-censo_SC	256	5
4	DS1-iris	150	4	21	DS3-idh_centro-oeste	465	3
5	DS1-maronna	200	2	22	DS3-idh_nordeste	1.780	3
6	DS1-outliers	150	2	23	DS3-idh_norte	449	3
7	DS1-ruspini	75	2	24	DS3-idh_sudeste	1.668	3
8	DS1-spherical_4d3c	400	3	25	DS3-idh_Sul	1.188	3
9	DS1-synthetic_control	600	51	26	DS3-pib_Ceara	164	4
10	DS1-wine	178	13	27	DS3-pib_ES	71	4
11	DS1-yeast	1.484	7	28	DS3-pib_MG	122	4
12	DS2-Gamma400	500	3	29	DS3-pib_PR	369	4
13	DS2-Normal300	300	2	30	DS3-pib_TO	117	4
14	DS2-Uniform400	400	2	31	DS4-Census	1.080	13
15	DS2-Uniform700	700	2	32	DS4-EIA	4.092	10
16	DS3-censo_MA	190	5	33	DS4-Tarragona	834	13
17	DS3-censo_MS	73	5	34	DS4-testdata	4.580	15

Experimento preliminar envolvendo as instâncias 1, 7, 31 e 33 foi realizado com o objetivo de calibrar os parâmetros do BRKGA, quais sejam: tamanho da população (p), tamanho do conjunto elite (p_e), número de soluções mutantes (p_m), probabilidade de seleção de genes do cromossomo elite no cruzamento (ρ_e) e total de gerações (g). O experimento de calibração abrangeu todas as combinações entre os valores dos parâmetros do BRKGA, bem como o nível de agregação (c). Os valores considerados para cada parâmetro são apresentados na Tabela 2.

Tabela 2 - Parâmetros

parâmetro	valores	parâmetro	valores
p	50, 75, 100	ρ_e	0.7, 0.8, 0.9
p_e	0.10, 0.15, 0.20, 0.25	g	100, 200, 300, 400, 500
p_m	0.10, 0.15, 0.20, 0.25	c	3, 4, 5, 10

Para cada combinação, 10 réplicas foram executadas, totalizando 28.800 execuções para cada instância. Por restrições de espaço, os resultados do experimento de calibração de parâmetros não foram apresentados em detalhes. Resumidamente, medidas de perda de informação e risco de violação de sigilo foram calculadas a partir dos resultados obtidos em cada uma das execuções. Testes de hipótese foram conduzidos a fim de identificar diferenças estatisticamente significativas no que diz respeito à redução da perda de informação e do risco de violação de sigilo. Análise de convergência também foi realizada para definir o número de gerações. A configuração identificada como de melhor desempenho contou com os seguintes valores de parâmetros: p = 100, p_e = 0.1, p_m = 0.15, ρ_e = 0.9, g = 200.

O método MABRKGA foi implementado em linguagem R. Aproveitando os recursos da arquitetura multicore do computador, e tendo disponível os pacotes `doParallel` e `foreach` do R, que agregam várias funções de paralelismo, o cálculo do ajuste foi paralelizado.

Os demais métodos já se encontram disponíveis em tal linguagem através do pacote `sdcMicro` [Templ 2007]. Em clustPPPCA, um método de agrupamento é aplicado antes de realizar a microagregação. O algoritmo utilizado para esta tarefa foi o CLARA [Kaufman e Rousseeuw 1986], adotado como método padrão pelo autor[2]. O número de grupos na etapa preliminar de agrupamento variou no intervalo [3, 8] e a melhor solução obtida foi considerada como solução final do método.

A qualidade do conjunto resultante após aplicação de um método de microagregação deve ser avaliada segundo a perda de informação e o risco de violação de sigilo. Neste sentido, três medidas foram então utilizadas, quais sejam: IL1 e IL2 para avaliar a primeira dimensão, relacionada à utilidade do conjunto protegido; e DLD para estimar o risco de violação de sigilo.

[2] Templ (2007) propõe o uso do algoritmo Mclust [Fraley e Raftery 1998], entretanto tal método não se encontra disponível no CRAN (do inglês, *Comprehensive R Archive Network*). A função `microaggregation` adota o algoritmo CLARA como padrão na execução de qualquer método que requeira etapa preliminar de agrupamento.

A medida IL1, Equação (5), representa a perda de informação normalizada. Está relacionada com o grau de homogeneidade dos grupos obtidos no processo de microagregação e é muito utilizada na literatura [Fayyoumi e Oommen 2010]. A medida IL2 [Mateo-Sanz et al. 2004], Equação (6), tem como objetivo quantificar o resultado da transformação do conjunto de dados original T no conjunto protegido T', ou seja, o impacto da transformação dos valores observados. No que diz respeito ao risco de violação do sigilo, *Linkage Disclosure* (LD) é o mecanismo padrão usado para medir o risco de violação de sigilo de um método de SDC [Domingo-Ferrer e Torra 2001]. Uma medida usada para quantificar o LD é o *Distance-based Linkage Disclosure* (DLD) [Herranz et al. 2012], descrito pela Equação (7). O resultado produzido a partir do cálculo do DLD possibilita quantificar a distância entre o conjunto de dados original padronizado T_N e o conjunto de dados protegido padronizado T'_N em termos do número de objetos com correspondência direta. Isto é, contabiliza o número de objetos $x'_j \in T'_N$ cujo objeto mais próximo em T_N seja o objeto x_j.

$$IL2 = \frac{1}{fn}\sum_{l=1}^{f}\sum_{j=1}^{n}\frac{\left|x_{jl}-x'_{jl}\right|}{\sqrt{2}S_l} \quad (6)$$

$$DLD = \frac{\#\left\{x'_j \mid j=\arg\min_m\left(d\left(x'_j,x_m\right)\right), x'_j\in T'_N, x_m\in T_N\right\}}{n}\times 100 \quad (7)$$

A Tabela 3 apresenta o *gap*[3], em termos de IL1 e IL2, para a solução vencedora de cada instância, dentre os cinco métodos. O nível de agregação considerado foi c = 3. A coluna DLD traz a medida de risco de violação de sigilo para a solução de menor IL1.

Tabela 3 - Resultados para nível de agregação c = 3

#	MABRKGA			MDAV			RMDM			PPPCA			clustPPPCA		
	$IL1_{gap}$	$IL2_{gap}$	DLD	$IL1_{gap}$	$IL2_{gap}$	DLD	$IL1_{gap}$	$IL2_{gap}$	DLD	$IL1_{gap}$	$IL2_{gap}$	DLD	$IL1_{gap}$	$IL2_{gap}$	DLD
1	**0,0**	**0,0**	25,0	92,1	1,7	29,0	78,3	3,9	30,0	661,0	220,5	20,0	295,8	100,0	21,0
2	3,2	**0,0**	24,0	**0,0**	**0,0**	31,0	2,9	0,9	30,8	25,1	15,8	18,5	25,0	15,9	19,0
3	46,1	17,8	25,3	0,2	0,6	32,1	**0,0**	**0,0**	31,9	11212,0	761,6	3,4	1007,5	198,0	15,1
4	**0,0**	3,2	25,3	34,5	**0,0**	28,0	33,8	5,6	28,0	120,4	70,6	19,3	212,4	59,2	24,7
5	**0,0**	**0,0**	26,0	47,5	4,9	29,5	20,7	1,7	31,5	1741,2	333,8	10,5	313,9	77,2	21,5
6	**0,0**	**0,0**	27,5	30,9	7,2	32,1	31,0	8,2	32,1	161,5	145,7	17,6	107,3	58,3	22,1
7	**0,0**	**0,0**	26,7	163,3	18,3	32,0	77,5	7,1	32,0	356,1	98,7	25,3	313,8	25,1	32,0
8	**0,0**	5,6	23,8	2,3	**0,0**	31,8	27,6	7,2	28,8	350,4	129,5	10,0	237,1	81,0	15,0
9	6,0	4,1	25,2	**0,0**	**0,0**	31,3	1,5	0,6	31,2	34,9	21,7	18,2	33,8	20,8	18,7
10	638,0	24,2	14,6	2502,1	1,6	5,1	2525,7	**0,0**	5,6	**0,0**	63,0	29,2	845,6	63,9	27,5
11	19,6	22,0	24,9	**0,0**	**0,0**	30,5	2,8	0,5	30,3	635,1	195,8	4,0	288,2	120,3	7,9
12	16,4	20,2	24,4	3,6	1,8	30,6	**0,0**	**0,0**	31,2	1088,2	321,6	6,0	222,2	112,5	13,6
13	33,1	12,7	25,0	1,3	1,7	33,0	**0,0**	**0,0**	32,7	2011,2	311,0	8,7	550,0	105,7	20,7
14	28,6	11,5	24,5	10,0	4,6	31,8	**0,0**	**0,0**	32,3	5352,7	495,9	8,0	938,9	149,7	21,0
15	52,4	19,1	25,6	**0,0**	**0,0**	32,3	4,3	1,9	32,0	9283,5	656,5	5,9	1372,7	206,5	16,4
16	**0,0**	13,0	24,2	13,3	0,8	25,8	6,4	**0,0**	28,9	218,9	85,9	9,5	64,7	42,2	18,9
17	**0,0**	0,3	27,4	1,1	**0,0**	27,4	21,5	4,0	30,1	124,2	54,6	19,2	75,1	19,6	24,7
18	**0,0**	5,6	25,0	26,7	4,8	25,8	14,4	**0,0**	27,3	139,8	59,6	16,4	71,7	34,8	20,3
19	14,8	5,9	27,3	**0,0**	**0,0**	31,8	21,2	10,3	30,7	215,3	82,7	12,5	104,3	37,4	27,3
20	4,4	8,8	24,6	5,3	3,2	29,3	**0,0**	**0,0**	30,5	223,1	96,1	13,7	104,6	54,7	19,1
21	28,0	14,1	25,4	4,0	1,3	31,6	**0,0**	**0,0**	31,6	575,1	195,4	8,2	279,7	101,7	15,5
22	49,9	21,6	24,9	1,3	**0,0**	31,2	**0,0**	0,3	31,0	1856,4	369,6	3,8	860,6	203,3	6,9
23	31,4	20,3	24,5	3,7	4,0	30,5	**0,0**	**0,0**	31,0	572,1	174,1	7,8	304,6	104,9	15,1
24	44,8	22,4	24,5	**0,0**	**0,0**	32,0	0,5	1,3	31,6	1323,6	322,9	4,9	731,7	193,0	7,3
25	51,1	19,5	24,7	1,9	**0,0**	30,6	**0,0**	0,2	30,9	1289,4	292,8	4,9	608,7	162,6	10,0
26	**0,0**	10,2	24,4	39,0	1,8	27,4	54,0	**0,0**	25,0	346,5	130,9	11,6	103,3	55,7	20,1
27	**0,0**	8,0	28,2	54,4	**0,0**	25,4	58,8	1,3	26,8	126,2	71,1	12,7	41,2	32,9	25,4
28	8,1	6,9	26,2	7,5	1,2	32,0	**0,0**	**0,0**	30,3	330,3	119,1	10,7	115,9	47,2	18,9
29	14,7	19,6	24,9	**0,0**	**0,0**	29,3	1,9	0,1	30,6	399,2	159,0	10,0	202,1	83,6	16,0
30	**0,0**	9,3	26,5	29,2	**0,0**	28,2	33,4	3,6	26,5	204,2	102,9	13,7	112,5	51,7	25,6
31	**0,0**	43,2	24,2	316,7	2,8	14,9	241,9	**0,0**	15,8	570,1	268,9	6,3	248,3	125,6	10,2
32	**0,0**	85,2	24,0	24,7	**0,0**	27,0	19,1	2,2	26,4	708,7	445,4	6,3	412,0	391,8	6,3
33	**0,0**	23,1	23,0	68,3	**0,0**	16,1	57,8	1,7	16,7	3,0	59,6	8,8	46,1	53,4	9,4
34	**0,0**	224,0	24,1	6051,4	0,1	0,8	6057,9	**0,0**	0,8	12929,3	244,5	2,0	1674,1	223,8	5,5
\	16	4	-	7	16	-	10	14	-	1	0	-	0	0	-

[3] O *gap* de IL1 é obtido pela seguinte expressão: $IL1_{gap} = \frac{IL1-IL1_{\text{solução vencedora}}}{IL1_{\text{solução vencedora}}}\times 100$. O cálculo é análogo para IL2.

https://proceedings.science/p/85390?lang=pt-br

As soluções vencedoras (*gap* = 0) foram identificadas em negrito e a última linha da Tabela 3 mostra o total de soluções vencedoras para cada método, segundo as duas medidas de perda de informação adotadas. Tendo como critério a medida de perda de informação IL1, o método MABRKGA obteve 16 (47%) soluções vencedoras, seguido pelo RMDM, com 10 (29%). Segundo esse critério, o método MDAV obteve 7 (21%) soluções vencedoras e o método PPPCA, 1 (3%) solução vencedora. Tomando IL2 como critério, o método MDAV obteve 16 (47%) soluções vencedoras, seguido pelo RMDM com 14 (41%) soluções vencedoras. O método proposto, MABRKGA, obteve 4 (12%) soluções vencedoras segundo esse critério. O método clustPPPCA não produziu soluções vencedoras em ambos os critérios de perda de informação considerados. O método PPPCA não produziu soluções vencedoras segundo o critério IL2.

Duas instâncias se destacaram nos resultados apresentados na Tabela 3, quais sejam: DS1-wine (10) e DS4-testdata (34). Os métodos MDAV e RMDM tiveram péssima performance em termos de IL1, em ambos os casos, porém produziram os melhores resultados em termos de IL2. Templ (2006) aponta limitações da medida IL2 em indicar a preservação de estatísticas multivariadas e as duas instâncias apresentam condição particular no que diz respeito à escala das variáveis. Esse fato será objeto de investigação futura.

No que diz respeito ao risco de violação de sigilo, medido por DLD, o método proposto, MABRKGA, apresentou o melhor desempenho geral. É esperado que, quanto menor a perda de informação, maior o risco de violação de sigilo, e vice-versa. Entretanto, na maioria dos casos em que o método alcançou a solução vencedora em termos de IL1, também apresentou o melhor resultado em termos de DLD (sublinhados na Tabela 3), em comparação com as soluções obtidas pelos métodos MDAV e RMDM.

As Figuras 2 e 3 apresentam os resultados em termos de IL1 e DLD, respectivamente, para os métodos MABRKGA e MDAV, considerando c = {3, 4, 5, 10}. As distribuições dos valores de IL1 para as soluções obtidas pelo método MABRKGA apresentaram menor dispersão e mediana inferior, em comparação com os resultados do método MDAV. Comportamento análogo foi observado em relação aos valores de DLD. A menor dispersão na distribuição de tais medidas, considerando que as instâncias de teste utilizadas possuem características bem distintas, sugere que o método MABRKGA é mais robusto, pois apresentou desempenho mais estável.

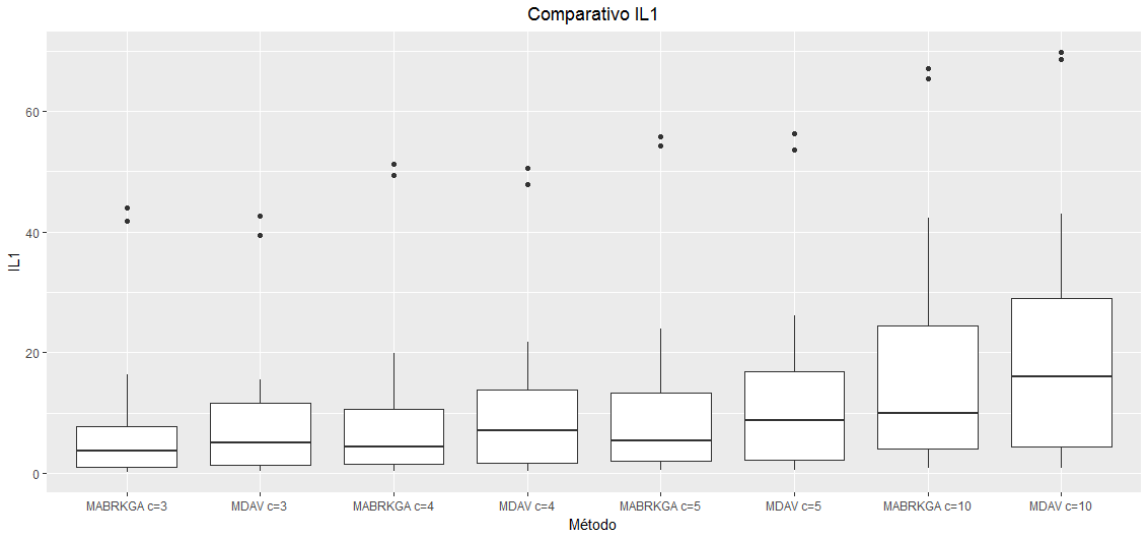


Figura 2 - Comparativo IL1, métodos MABRKGA e MDAV

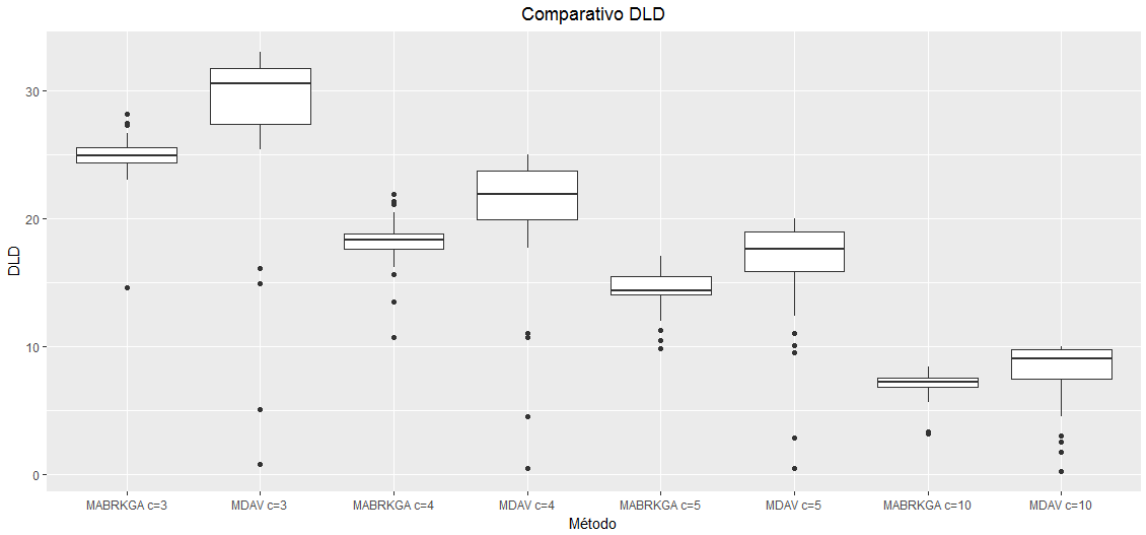


Figura 3 - Comparativo DLD, métodos MABRKGA e MDAV

A Tabela 4 apresenta os valores utilizados na construção das Figuras 2 e 3 a fim de permitir comparação com o MDAV, bem como com outros métodos.

Tabela 4 - Resultados para níveis de agregação c = {3, 4, 5, 10}

#	MABRKGA								MDAV							
	c=3		c=4		c=5		c=10		c=3		c=4		c=5		c=10	
	IL1	DLD	IL1	DLD	IL1	DLD	IL1	DLD	IL1	DLD	IL1	DLD	IL1	DLD	IL1	DLD
1	0,27	25,00	0,37	18,50	0,48	15,50	1,03	7,50	0,53	29,00	0,45	20,50	0,60	19,50	0,94	9,50
2	44,00	24,00	51,26	13,50	55,75	9,83	67,11	3,17	42,65	31,00	50,57	17,67	56,30	11,00	69,75	2,50
3	0,42	25,33	0,61	18,44	0,81	14,67	1,79	7,00	0,29	32,11	0,41	23,89	0,57	19,44	1,24	9,89
4	1,54	25,33	2,11	18,00	2,65	16,00	4,54	7,33	2,07	28,00	2,97	21,33	4,74	17,33	8,35	8,67
5	0,89	26,00	1,45	18,50	1,79	15,50	4,03	8,00	1,31	29,50	1,72	24,00	1,95	20,00	4,27	9,50
6	9,05	27,48	12,99	19,85	15,65	15,27	25,40	8,40	11,84	32,06	14,80	24,43	17,00	19,85	24,02	9,92
7	0,96	26,67	1,48	21,33	1,93	16,00	4,06	8,00	2,53	32,00	4,76	21,33	3,82	17,33	16,94	8,00
8	0,38	23,75	0,51	18,25	0,67	14,00	1,23	7,25	0,39	31,75	0,42	22,50	0,53	18,25	0,98	9,75
9	41,89	25,17	49,47	16,17	54,21	10,50	65,43	3,33	39,51	31,33	47,92	20,00	53,66	12,33	68,56	3,00
10	4,08	10,67	4,35	14,61	5,57	11,24	10,35	7,30	15,35	5,06	21,75	4,49	21,97	2,81	42,57	1,69
11	6,22	24,93	8,33	17,52	10,29	12,87	17,29	6,13	5,20	30,53	7,16	22,17	8,61	16,98	15,19	8,83
12	3,39	24,40	4,54	18,80	5,39	14,40	9,62	7,20	3,01	30,60	3,78	22,60	4,91	18,60	9,89	9,80
13	2,38	25,00	3,29	18,67	4,43	15,00	9,09	7,00	1,81	33,00	2,85	25,00	3,70	19,00	8,22	9,67
14	0,72	24,50	1,07	18,75	1,48	14,50	3,28	7,25	0,62	31,75	0,93	23,75	1,20	19,00	2,78	10,00
15	0,49	25,57	0,73	18,00	0,95	14,14	2,17	7,14	0,32	32,29	0,51	24,14	0,63	19,57	1,45	10,00
16	7,70	24,21	10,56	17,89	12,54	14,74	19,72	6,84	8,72	25,79	12,10	19,47	15,07	13,68	22,61	7,37
17	12,00	27,40	16,17	21,92	19,66	16,44	31,49	8,22	12,13	27,40	18,53	21,92	22,99	17,81	38,83	6,85
18	7,75	25,00	10,41	18,75	13,60	14,06	21,98	7,03	9,82	25,78	11,93	20,31	14,27	17,19	25,05	8,59
19	16,43	27,27	19,85	20,45	23,88	17,05	42,40	7,95	14,31	31,82	18,49	23,86	26,10	18,18	43,01	9,09
20	7,30	24,61	9,75	18,75	11,28	14,06	18,04	7,81	7,37	29,30	9,85	21,88	11,92	18,36	19,98	8,98
21	4,02	25,38	5,48	17,63	6,78	14,19	12,32	7,31	3,26	31,61	4,55	23,44	5,81	18,28	10,15	9,46
22	1,76	24,94	2,45	18,26	3,09	14,21	5,86	6,29	1,19	31,24	1,73	23,82	2,19	19,27	4,17	9,72
23	3,00	24,50	4,17	18,04	5,14	14,25	9,10	7,57	2,37	30,51	3,33	24,05	4,32	18,71	7,51	9,58
24	1,25	24,46	1,77	17,57	2,22	14,03	4,06	6,53	0,86	31,95	1,27	23,20	1,61	18,59	2,88	9,77
25	2,08	24,66	2,95	17,76	3,66	14,48	7,00	7,07	1,41	30,56	1,94	23,57	2,49	19,36	4,67	9,85
26	6,86	24,39	9,50	19,51	10,95	15,85	20,30	7,32	9,53	27,44	12,88	20,12	16,09	14,02	25,85	9,15
27	8,15	28,17	12,22	21,13	16,14	15,49	32,04	7,04	12,59	25,35	17,76	18,31	19,82	16,90	37,51	8,45
28	8,84	26,23	14,18	20,49	18,15	15,57	30,39	8,20	8,79	31,97	13,09	20,49	13,79	17,21	24,22	9,84
29	5,58	24,93	7,50	18,70	8,87	14,36	15,63	7,05	4,86	29,27	7,17	22,22	8,81	17,34	14,08	8,13
30	8,75	26,50	11,30	18,80	14,22	16,24	25,70	7,69	11,31	28,21	13,29	19,66	16,28	17,09	31,65	7,69
31	2,29	24,17	2,98	17,50	3,67	13,52	6,33	6,76	9,54	14,91	11,39	11,02	13,70	9,54	22,75	4,54
32	0,33	24,00	0,47	17,33	0,65	13,71	1,40	6,50	0,42	26,98	0,56	19,84	1,07	15,49	2,16	7,21
33	7,85	23,02	10,58	15,59	12,44	11,99	25,42	5,64	13,21	16,07	14,05	10,67	19,45	10,07	30,16	5,16
34	0,25	24,08	0,36	17,42	0,46	13,84	0,90	6,55	15,59	0,79	19,99	0,44	23,81	0,41	34,29	0,24

6. Conclusões e trabalhos futuros

O método MABRKGA se mostrou competitivo e superou os demais métodos considerados, no que diz respeito ao número de soluções vencedoras, segundo o critério de perda de informação IL1. Em especial, no que diz respeito aos conjuntos de dados típicos da literatura de SDC (31 a 34), o MABRKGA produziu as melhores soluções em termos de IL1. Além disso, embora seja esperado que, quanto menor a perda de informação, maior o risco de violação de sigilo, o método proposto apresentou o melhor resultado em termos de DLD na maioria dos casos em que alcançou a solução vencedora, em comparação com o DLD das soluções obtidas pelos métodos MDAV e RMDM.

O método MDAV alcançou o maior número de soluções vencedoras quando adotado o critério de perda de informação IL2. O método RMDM, embora não tenha obtido o maior número de soluções vencedoras em nenhum dos dois critérios utilizados, foi o que apresentou melhor desempenho se considerados ambos os critérios. Quando avaliadas segundo o critério de perda de informação IL2, apenas 4 (12%) das soluções obtidas pelo MABRKGA foram vencedoras. Sob esse critério, o método proposto foi superado pelo MDAV e RMDM. A diferença de desempenho do MABRKGA entre os dois critérios de perda de informação adotados pode ser justificada pela função utilizada para cálculo do ajuste e, portanto, escolha das melhores soluções.

Ainda no que diz respeito aos critérios utilizados, segundo Templ (2006), IL2 avalia apenas estatísticas univariadas e, portanto, pode não indicar corretamente se estatísticas multivariadas foram preservadas. Outra vantagem do método proposto é que este pode ser facilmente adaptado para obter soluções segundo outra função objetivo.

Em relação ao tempo de execução, embora valores não tenham sido apresentados na Seção 5, o MABRKGA teve eficiência muito inferior aos demais métodos, especialmente quando comparado ao MDAV. Os métodos disponíveis no pacote sdcMicro contam com implementações extremamente eficientes, aspecto que foi pouco explorado na implementação do método proposto. O principal recurso utilizado foi paralelização do cálculo do ajuste, etapa mais custosa do método. Outras estratégias de paralelização, bem como o uso de estruturas de dados mais eficientes podem ser consideradas para que o MABRKGA se torne mais competitivo em termos de tempo de execução.

Outro aspecto que requer investigação mais aprofundada é o método de inicialização da população. Embora a inicialização através do PCVE tenha apesentado resultados competitivos, o custo computacional para obter soluções para esse problema em conjuntos de dados com muitos objetos pode representar um problema. Uma possível abordagem é o uso da fase de particionamento do método MDAV para tal fim. O método oferece soluções de boa qualidade, o que pode acelerar a convergência do MABRKGA, e conta com implementação extremamente eficiente, podendo contribuir também no ganho de eficiência global. Outra possibilidade é adotar algoritmos de agrupamento particionais não determinísticos, tais como k-means e CLARA.

Ainda em relação a possíveis trabalhos futuros, métodos mais sofisticados podem ser adotados para realizar a calibração dos parâmetros do BRKGA, tal como metodologia apresentada em Moro (2017).

Referências

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

Domingo-Ferrer, J. e Mateo-Sanz, J. (2002). Practical data-oriented microaggregation for statistical disclosure control. *IEEE Trans. on Knowledge and Data Engineering*, 14(1):189-201.

Domingo-Ferrer, J. e Torra, V. (2001). A quantitative comparison of disclosure control methods for microdata. Confidentiality, disclosure, and data access: Theory and practical applications for statistical agencies, Elsevier.

Domingo-Ferrer, J. e Torra, V. (2003). On the connections between statistical disclosure control for microdata and some artificial intelligence tools. Information Sciences, 151:153-170.

Domingo-Ferrer, J. e Torra, V. (2005). Ordinal, continuous and heterogenerous k-anonymity through microaggregation. Data Mining and Knowledge Discovery 11(2), 195–212.
Domingo-Ferrer, J., Martínez-Ballesté, A., Mateo-Sanz, J. e Sebé, F. (2006). Efficient Multivariate Data-oriented Microaggregation. The VLDB Journal, 15(4):355-369.
Domingo-Ferrer, J., Mateo-Sanz, J., e Torra, V. (2001). Comparing SDC Methods for Microdata on the Basis of Information Loss and Disclosure Risk. Proceedings of ETK-NTTS, 2:807-826.
Elliot, M., Hundepool, A., Nordholt, E.S., Tambay, J-L. e Wende, T. (2005). Glossary on SDC.
Fayyoumi, E. e Oommen, B. (2010). A survey on statistical disclosure control and micro-aggregation techniques for secure statistical databases. Software: Practice and Experience, 40(12), 1161-1188.
Fraley, C. e Raftery, A. E. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. The Computer Journal, 41(8):578–588.
Gonçalves, J. e Resende, M. (2011). Biased Random-key Genetic Algorithms for Combinatorial Optimization. Journal of Heuristics, 17(5):487-525.
Hansen, S. e Mukherjee, S. (2003). A Polynomial Algorithm for Optimal Univariate Microaggregation. IEEE Trans. on Know. and Data Eng., 15(4), 1043-1044.
Heaton, W. (2011). New Record Ordering Heuristics for Multivariate Microaggregation. Tese (doutorado), Nova Southeastern University.
Herranz, J., Nin, J. e Solé, M. (2012). Kd-trees and the Real Disclosure Risks of Large Sta-tistical Databases. Information Fusion, 13(4), 260-273.
Jolliffe, I. T. (2002). Principal Component Analysis, Springer Series in Statistics, 2nd ed., Springer, NY.
Kaufman, L. e Rousseeuw, P. J. (1986). Clustering Large Data Sets. E. S. Gelsema and L. N. Kanal (Eds.), Pattern Recognition in Pratice II. Elsevier, North Holland.
Laszlo, M. and Mukherjee, S. (2005). Minimum spanning tree partitioning algorithm for mi-croaggregation. IEEE Trans. on Know. and Data Eng., 17(7):902-911.
Li, G. e Chen, Z. (1985). Projection-pursuit approach to robust dispersion matrices and principal components: primary theory and monte carlo. J. Amer. Statist. Ass., 80:759-766.
Liu, H. e Motoda, H. (1998). Feature Selection for Knowledge Discovery and Data Mining. Kluwer Academic.
Maronna, R. A. e Zamar, R. H. (2002). Robust multivariate estimates for high dimensional datasets. Technometrics, 44:307-317.
Mateo-Sanz, J. M., Sebe, F., e Domingo-Ferrer, J. (2004). Outlier protection in continuous microdata masking. Lecture Notes in Comp. Sci. Privacy in Statistical Databases, Springer Verlag, 3050:201-215.
Mitchell, M. (1998). An introduction to genetic algorithms. MIT Press.
Moro, M. A. (2017). Meta-heurísticas GRASP e BRKGA aplicadas ao Problema da Diversidade Máxima. Dissertação (mestrado), Universidade Estadual de Campinas - UNICAMP.
Oganian, A. e Domingo-Ferrer, J. (2001). On the complexity of optimal microaggregation for statistical disclosure control. Statistical Journal of the UNECE, 18(4):345-353.
Rosenkrantz, D., Stearns, R. e Lewis, II, P. (1977). An Analysis of Several Heuristics for the Traveling Salesman Problem. Journal on Computing, 6(3):563-581.
Schmid, M. (2009). The effect of single-axis sorting on the estimation of a linear regression. Journal of Official Statistics, 25(4):529-548.
Solanas, A. (2008). Privacy Protection with Genetic Algorithms. Success in Evolutionary Computation, Springer Berlin Heidelberg.
Solanas, A., Martínez-Ballesté, A., Mateo-Sanz, J. e Domingo-Ferrer, J. (2006). Multivariate microaggregation based genetic algorithms. 3rd International IEEE Conf. Intelligent Sys., London, 65-70.
Templ, M. (2006) Software development for SDC in R. In: Domingo-Ferrer, J., Franconi, L. (eds.) PSD 2006. LNCS, 4302:347-359. Springer, Heidelberg.
Templ, M. (2007). sdcMicro: A new flexible R-package for the generation of anonymised microdata - design issues and new methods. In: Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, Monographs of Official Statistics.
Templ, M., Meindl, B. e Kowarik, A. (2018). Introduction to Statistical Disclosure Control (SDC). International Household Survey Network.
UNSTATS. (2017a). Report of the Global Working Group on Big Data for Official Statistics. Documents of the 49th session of the UN Statistical Commission.
UNSTATS. (2017b). Report of the Secretary-General on open data. Documents of the 49th session of the UN Statistical Commission.