Proposta de método de microagregação para controle estatístico de sigilo através de algoritmo genético de chaves aleatórias viciadas

Augusto Fadel[1]✉, Luiz S. Ochi[2], Gustavo S. Semaan[3], José André de M. Brito[4]

[1]IBGE, Rio de Janeiro/RJ, Brasil
[2]IC/UFF, Niterói/RJ, Brasil
[3]INFES/UFF, Santo Antônio de Pádua/RJ, Brasil
[4]ENCE/IBGE, Rio de Janeiro/RJ, Brasil
augustofadel@gmail.com

Abstract. A evolução tecnológica observada nos últimos tempos, aliada à crescente popularização de sistemas e recursos computacionais, tem permitido que um substancial volume de informações seja captado, analisado e armazenado. Embora muito úteis para o meio científico, a disseminação de tais fontes de dados envolve o aspecto ético do risco de violação do sigilo do indivíduo. Neste sentido, a microagregação, técnicas de controle estatístico de sigilo, pode contribuir para mitigar este risco. O presente trabalho apresenta proposta de um novo método de microagregação, considerando o estudo da metaheurística algoritmo genético de chaves aleatórias. No final do trabalho é apresentado um conjunto de resultados computacionais promissores, sendo esses resultados referentes à aplicação do novo método proposto e de outros métodos da literatura.

Keywords: Microagregação · Sigilo · Algoritmo Genético · BRKGA

1 Introdução

A produção e a divulgação de dados públicos já foi uma atividade exclusiva de alguns poucos entes especializado nesse fim. A evolução tecnológica ocorrida especialmente nas últimas duas décadas permitiu a captação e o armazenamento de grandes volumes de dados por diversos agentes, com diferentes objetivos. Além disso, favoreceu a democratização do acesso e uso (análise) desses dados. No entanto, os métodos adotados por tais agentes nem sempre garantem alguns aspectos importantes, tais como qualidade (e.g. métodos adotados para coleta e armazenamento utilizados), representatividade (no que diz respeito à população de interesse) e periodicidade (i.e. com que frequência e até quando a informação será disponibilizada).

Os Institutos Nacionais de Estatística (INE) são ainda o principal provedor de dados e informações, atendendo a necessidades dos mais diversos segmentos da sociedade civil, bem como dos órgãos das esferas governamentais. Os INE operam segundo os Princípios Fundamentais das Estatísticas Oficiais [36], estabelecidos pela Divi-

são Estatística da Organização das Nações Unidas (UNSTAT), garantindo amplamente os aspectos mencionados anteriormente, dentre outros.

Além dos aspectos relacionados à qualidade do dado, os Princípios Fundamentais estabelecem o conceito de confidencialidade, no qual o sigilo do informante deve ser garantido e as informações por ele fornecidas utilizadas exclusivamente para fins estatísticos, excluindo, portanto, o uso para fins fiscais, por exemplo.

A divulgação do dado tabulado e a aplicação de técnicas de controle estatístico de sigilo (do inglês, *Statistical Disclosure Control* - SDC) impedem que informações confidenciais sejam inadvertidamente reveladas [19]. Embora essa forma de divulgação atenda ao usuário comum, ou seja, o indivíduo que, por interesse pessoal ou profissional, deseja obter informações quantitativas sobre algum segmento da sociedade civil, é, no entanto, insuficiente para o usuário avançado. Por exemplo, os Analistas e membros da academia, ao desenvolverem um estudo, demandam dados com nível de detalhe muito maior [34] e desejam, em geral, obter os dados individualizados (microdados), tornando a questão do sigilo ainda mais delicada.

A fim de atender usuários avançados (como pesquisadores acadêmicos), cujos propósitos não violem o estabelecido nos Princípios Fundamentais, é necessário buscar meios alternativos de disseminação de dados. Neste sentido, uma possível abordagem é a aplicação de métodos de SDC para microdados, tais como adição de ruído [2], em que, para cada atributo j, o vetor de valores observados x_j é substituído por $z_j=x_j+\epsilon_j$, sendo $\epsilon_j \sim N(0,\sigma^2_{\epsilon_j})$; geração sintética de microdados [10], onde um novo conjunto é construído a partir de dados gerados aleatoriamente, porém preservando determinadas estatísticas do conjunto original, tais como média e variância; microagregação [5, 11], onde os objetos (e.g. indivíduos) são reunidos em grupos de tamanho mínimo k e os valores observados são substituídos pela média do grupo; e métodos híbridos [19], quando o conjunto final é composto por dados sintéticos e originais combinados.

Os primeiros métodos de SDC para microdados foram desenvolvidos na década de 1980 [10]. Nesse momento, no entanto, o objetivo era apenas a proteção do sigilo, fazendo com que os resultados obtidos nas primeiras abordagens propostas fossem insatisfatórios no que diz respeito à preservação da informação, inviabilizando a realização de análises e estudos bem sucedidos. Alguns métodos "ingênuos" falham não apenas na preservação da informação, mas, em determinadas condições, também na proteção do sigilo [37]. Em comparativo [6] envolvendo diversos métodos de SDC, a microagregação multivariada apresentou melhor desempenho no que diz respeito à redução da perda de informação e do risco de violação do sigilo.

No aspecto computacional, o problema de microagregação multivariado ótimo é NP-difícil [28], o que implica impossibilidade da aplicação de um método de enumeração exaustiva para resolvê-lo. Segundo Torra e Navarro-Arribas [35], a microagregação é baseada em métodos de agrupamento. Brito et al. [3] propuseram algoritmo para solução do problema dos k-medoids através do Algoritmo Genético de Chaves Aleatórias Viciadas [13], que obteve bom desempenho no que diz respeito à obtenção de soluções ótimas para o problema de agrupamento.

O problema de microagregação é abordado em duas etapas, sejam elas: particionamento e agregação. Os algoritmos disponíveis na literatura, de maneira geral, empregam métodos de agrupamento baseados em centroide na fase de particionamento.

A presença de valores extremos (outliers) pode influenciar severamente o valor médio de um grupo, i.e. o centroide, afetando, consequentemente, a alocação dos demais objetos. O medóide, por sua vez, é um objeto do conjunto de dados e, portanto, quando utilizado como protótipo do grupo é robusto à presença de valores extremos e ruídos, possibilitando a formação de grupos mais coesos e homogêneos.

Nesse contexto, o presente trabalho apresenta proposta de método de microagregação multivariado para atributos numéricos contínuos com grupos de tamanho variável, baseando-se, para isso, em um algoritmo de otimização, que foi aplicado ao problema de controle estatístico de sigilo. Além da introdução, esse artigo conta ainda com cinco outras seções. Na Seção 2 é apresentada uma descrição formal do problema e a Seção 3 traz uma revisão da literatura. A metodologia proposta é descrita na Seção 4 e os resultados do experimento computacional realizado são apresentados na Seção 5. Finalmente, a Seção 6 apresenta as conclusões obtidas e possíveis trabalhos futuros.

2 Descrição do problema

O problema de microagregação clássico pode ser definido em duas etapas [4], sejam elas: (i) Particionamento: dado um nível de agregação k, a base de dados composta por n registros é particionado 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, k objetos; (ii) Agregação: um operador de agregação (e.g. média, para variáveis numéricas) é utilizado para computar o centroide de cada grupo. Cada objeto é então substituído pelo centroide do grupo em que está alocado.

No que diz respeito à otimização, a solução ótima para o problema de microagregação é aquela na qual a máxima homogeneidade interna dos grupos é obtida [11]. O problema pode então ser formalizado como: em uma base de dados T, com p atributos numéricos contínuos e n objetos, c grupos devem ser construídos de forma a minimizar o erro quadrático total (do inglês, *sum of squared errors* - SSE), ou seja, minimizar o somatório dos quadrados das diferenças dos objetos e seus respectivos centroides, dado pela Equação (1).

$$\mathrm{SSE} = \sum_{i=1}^{c} \sum_{j=1}^{n_i} \sum_{l=1}^{p} \left(x_{ijl} - \bar{x}_{il}\right)^T \left(x_{ijl} - \bar{x}_{il}\right) \quad (1)$$

Onde x_{ijl} representa o valor do l-ésimo atributo para o j-ésimo objeto do grupo G_i, $\bar{x}_{il}$ é a média do atributo l para os objetos no grupo G_i $\left(\bar{x}_{il} = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ijl}\right)$, e n_i representa o número de objetos do grupo G_i. A fim de obter uma solução adequada, as seguintes restrições devem ser consideradas:

$$\sum_{i=1}^{c} n_i = n \quad (2)$$

$$k \leq n_i < 2k \; \forall \; i \in \{1, \ldots, c\} \quad (3)$$

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

Duas restrições (Equações 2 e 4) indicam que todo objeto deve ser alocado em um e somente um grupo. A restrição apresentada na Equação 3 limita o número de objetos em cada grupo. Segundo Domingo-Ferrer e Mateo-Sanz [5], o tamanho dos grupos (n_i) da solução ótima é menor do que 2k, para todo $i \in \{1, \ldots, c\}$, sendo k o nível de agregação, dado de entrada do problema.

3 Revisão da literatura

Os métodos de SDC podem ser classificados de acordo com o tipo de aplicação que o usuário final dará ao dado. Mais especificamente, tais métodos são divididos em três categorias [35], são elas: orientados a computação (*computation-driven*), orientados a resultados (*result-driven*) e orientados a dados (*data-driven*).

Os métodos orientados a dados, adequados ao propósito de divulgação e disseminação de microdados consistem, basicamente, em modificar o conjunto de dados reduzindo a quantidade de informação nele disponível, a fim de garantir o sigilo. 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 microagregação consiste em um método de SDC perturbativo [4].

De acordo com Mortazavi e Jalili [26], os métodos de microagregação foram originalmente propostos para atributos numéricos, mas há variações que admitem outros tipos de variáveis [9]. Neste trabalho foram contemplados métodos para solução do problema de microagregação multivariado para atributos numéricos contínuos.

No que diz respeito ao problema de microagregação univariado, ou seja, quando apenas um atributo é considerado, Hansen e Mukherjee [16] apresentaram algoritmo, chamado MHM, produz o valor mínimo de SSE em tempo polinomial, com base no problema do caminho mínimo (*shortest path problem*). Essa abordagem consiste em construir um grafo acíclico direcionado onde os vértices representam os objetos do conjunto de dados. Cada vértice é então conectado aos 2k-1 vértices mais similares a ele, segundo métrica previamente definida, e cada aresta recebe como peso o SSE dos objetos associados aos seus vértices. Inicialmente, o algoritmo obtém o caminho mínimo passando por todos os vértices e, em seguida, particiona esse caminho a cada k vértices. Mortazavi et al. [25] propuseram método iterativo baseado no algoritmo MHM para o caso multivariado, chamado IMHM. O caso univariado é um problema pertencente à classe P, no entanto, Oganian e Domingo-Ferrer [28] mostraram que o problema de microagregação multivariado ótimo é NP-difícil, justificando assim a aplicação de heurísticas e metaheurísticas para resolvê-lo.

Segundo Solanas [32], o método da Distância Máxima para o Vetor Médio (do inglês, *Maximum Distance to Average Vector* - MDAV) [9] é o algoritmo de microagregação mais utilizado. Esse método consiste em calcular o centroide considerando todos os objetos do conjunto de dados e, em seguida, construir dois grupos de tamanho k, sendo um ao redor do objeto (r) com menor similaridade em relação ao centróide e outro ao redor do objeto com menor similaridade em relação ao objeto r. O centróide é então atualizado para os objetos restantes (não agrupado) e o processo se repete até que todos os objetos tenham sido alocados a algum grupo.

Laszlo e Mukherjee [21] propuseram método de Tamanho Fixo Baseado em Centroide (do inglês, *Centroid-based Fixed-size* - CBFS), similar ao MDAV, porém apenas o grupo ao redor do objeto r (mais distante do centroide) é construído em cada iteração. Tal alteração promove significativo ganho de eficiência, pois, no MDAV, o

número de operações por iteração para o cálculo das distâncias é $O(n^2)$, enquanto que apenas $O(n)$ operações são necessárias no CBFS. Em Solé et al. [33] é apresentada implementação eficiente do CBFS utilizando estrutura de dados *kd-tree* [1], com complexidade de caso médio $O(\log n)$, permitindo tratar grandes conjuntos de dados.

No método de seleção de grupos com base em minimização sequencial do SSE (do inglês, *Group Selection based on sequential Minimization of SSE* - GSMS) [29] cada objeto é um candidato a centroide, formando um grupo com os k-1 objetos mais próximos a ele. O processo é realizado em duas etapas. Na primeira são realizadas $\lfloor n/k \rfloor$ iterações nas quais é descartado o grupo que minimiza o SSE dos demais objetos, formando, ao final $\lfloor n/k \rfloor$ grupos de tamanho k. Na segunda etapa os $n-k\lfloor n/k \rfloor$ objetos restantes são alocados ao grupo mais próximo.

Heaton [17] apresentou o algoritmo MicTSP, uma variação do algoritmo MHM para o caso multivariado com base no problema do caixeiro viajante (do inglês, *Travelling Salesman Problem* - TSP), onde a ordenação dos objetos é executada segundo o próximo ponto em uma rota do TSP. O algoritmo MHM é então aplicado de acordo com a sequência de objetos estabelecida. De maneira similar, o algoritmo NPN-MHM (do inglês, *Nearest Point Next*) [4] define a sequência de objetos para aplicação do MHM de acordo com a abordagem do ponto mais próximo. Isto é, adotando como objeto inicial aquele mais distante do centroide do conjunto completo, o objeto seguinte da sequência será aquele que apresentar menor distância para o objeto anterior (o objeto inicial, na primeira iteração). Esse procedimento é aplicado até que todos os objetos do conjunto tenham sido ordenados.

O algoritmo FDM (do inglês, *Fast Data-oriented Microaggegation*) [27] consiste em aplicar o MHM em uma sequência de objetos definida por uma rota TSP. Entretanto, como a construção dessa sequência não depende do valor de k, os autores propuseram gerar soluções para diferentes valores do nível de agregação k em uma única execução. Em termos práticos, essa abordagem pode economizar bastante tempo de processamento, uma vez que, dado um algoritmo de microagregação, múltiplas execuções são realizadas dentro de um intervalo definido para k. A solução final é escolhida com base em métricas de perda de informação e risco de sigilo.

Alguns dos métodos apresentados produzem grupos de tamanhos iguais, com k objetos cada (quando o número de objetos n não é múltiplo de k, algum tipo de correção é empregada para alocar os objetos remanescentes). Mortazavi e Jalili [26] argumentam que tal restrição resulta em perda de desempenho no que diz respeito ao valor da função objetivo (minimização do SSE) em relação aos métodos que constroem os grupos obedecendo a restrição dada pela Equação (4), descrita na Seção 2. Quanto maior o valor da função objetivo para a solução considerada, maior é a perda de qualidade do conjunto resultante, no que diz respeito à preservação da informação contida no conjunto de dados original. De acordo com Li [22], entretanto, a abordagem de grupos de tamanho fixo é computacionalmente mais eficiente do que a abordagem que admite grupos de tamanho variável. Neste sentido, Fayyoumi e Oommen [11] afirmam que, embora os métodos envolvendo grupos de tamanho variável sejam marginalmente mais complexos do que aqueles que envolvem grupos de tamanho fixo, são menos propensos a comprometer o sigilo do microdado.

4 Metodologia proposta

Os métodos de microagregação são baseados em métodos utilizados para resolver problemas de agrupamento [35]. O problema de agrupamento clássico [14], por sua vez, consiste em particionar um conjunto de n objetos, descritos por p atributos, em c grupos de objetos similares, de acordo com uma métrica e uma função objetivo previamente definidas. A microagregação consiste em particionar um conjunto de n objetos, descrito por p atributos, em grupos de, ao menos, k objetos similares. Assumindo abordagem baseada em grupos de tamanho fixo, com k objetos cada, há correspondência direta entre os dois problemas, pois $c=\lfloor n/k \rfloor$. Entretanto, a microagregação com grupos de tamanhos variáveis pode ser vista como o problema de agrupamento automático [31], sendo o número de grupos não fixado previamente.

Muitos dos algoritmos de microagregação disponíveis na literatura aplicam, em sua fase de particionamento, algoritmos de agrupamento baseados em centroide. Entretanto, é sabido que a presença de valores extremos (*outliers*) pode influenciar severamente o valor médio de um grupo, i.e. o centroide, produzindo solução de baixa qualidade no que diz respeito à alocação adequada dos objetos. Segundo Han et al. [15], esse efeito é acentuado quando a função objetivo a ser minimizada no problema de agrupamento é definida em termos do SSE. Alternativamente ao centroide, temos o medóide, que é um objeto do conjunto de dados. Assim sendo, quando utilizado como protótipo do grupo**,** é mais robusto à presença de valores extremos e de ruídos.

A abordagem para o problema de agrupamento baseada em medóides foi formalizado como o problema dos k-medoids [20], ao qual técnicas de otimização podem ser aplicadas. Em Brito et al. [3] o problema dos k-medoids foi abordado utilizando conceitos da metaheurística algoritmo genético de chaves aleatórias viciadas (do inglês, *Biased Random Key Genetic Algorithm* – BRKGA) [13], com reconexão por caminhos (*path relinking*) [12]. No experimento computacional conduzido pelos autores, envolvendo 30 instâncias, o algoritmo proposto alcançou o ótimo global em aproximadamente 95% dos casos. Embora outros métodos de microagregação apliquem metaheurísticas, não foi identificada nenhuma abordagem baseada em medóides. Posto isso, o presente trabalho propõe um novo método de microagregação multivariado com grupos de tamanho variável**,** através da resolução do problema dos k-medoids e implementação baseada no BRKGA.

4.1 Formulação

A etapa de particionamento foi abordada através do problema dos k-medoids, onde inicialmente, para um conjunto de dados T, formado por n objetos descritos por p atributos, c objetos, denotados medóides ou objetos representativos [20], devem ser selecionados. Em seguida, os n-c objetos de T restantes são associados ao medóide mais próximo, segundo uma métrica de distância, formando então c grupos. Quanto mais homogêneos os grupos formados, maior é a qualidade da solução obtida e, portanto, os medóides devem ser escolhidos de forma a minimizar, em cada grupo, a média das distâncias entre o medóide e os demais objetos a ele alocados. A função objetivo a ser minimizada é dada pela Equação (5)

$$\sum_{i=1}^{c} \frac{1}{n_i} \sum_{j=1}^{n_i} d(x_j, m_i) \quad (5)$$

Onde n_i é o número de objetos alocados no grupo i, m_i é o medóide do grupo i e $d(x_j, m_i)$ é a distância euclidiana de cada objeto x_j, pertencente ao grupo i, e seu respectivo medóide, m_i. Além disso, foram consideradas as restrições dadas pelas Equações (2), (3) e (4), descritas na Seção 2.

4.2 Algoritmo

O problema de otimização foi abordado através de uma implementação do BRKGA. Assim sendo, cada cromossomo (solução) é definido como um vetor v com n posições (genes), correspondentes aos objetos de T. As posições de v são preenchidas com valores reais gerados segundo uma distribuição uniforme no intervalo $[0,1]$. Uma vez gerado v, o decodificador atribui a um vetor w os elementos de v ordenados crescentemente. Os c primeiros valores de w são então pesquisados em v e as posições retornadas correspondem aos objetos escolhidos como medóides, m.

Assumindo exemplo hipotético de um conjunto de dados com 10 objetos, a ser particionado em 3 grupos, a decodificação de uma solução é representada abaixo.

$$v = \{0.918, \mathbf{0.114}, \mathbf{0.219}, 0.397, 0.981, \mathbf{0.245}, 0.483, 0.546, 0.504, 0.898\}$$
$$w = \{\mathbf{0.114}, \mathbf{0.219}, \mathbf{0.245}, 0.397, 0.483, 0.504, 0.546, 0.898, 0.918, 0.981\}$$
$$m = \{2, 3, 6\}$$

Uma população inicial, constituída por p vetores de chaves aleatórias, ou cromossomos, é decodificada e a função objetivo, definida pela Equação (5), é computada para as soluções obtidas a partir de cada conjunto de medóides m_i, onde $i \in \{1, \ldots, p\}$. A população é então particionada em dois conjuntos, segundo os valores obtidos para a função objetivo. Um conjunto elite composto das p_e melhores soluções, ou seja, com menor valor da função objetivo (uma vez que o problema em questão é um problema de minimização), e um conjunto não elite, formado pelas $p\text{-}p_e$ demais soluções, sendo $p_e < p\text{-}p_e$.

Uma nova geração (g+1) de cromossomos é então produzida, a fim de atualizar a população. A partir da segunda geração, cada população é composta por três grupos de soluções, sejam eles: (i) C_{elite}, com as p_e soluções elite obtidas na geração anterior, g, integralmente replicadas para a geração seguinte, g+1; (ii) $C_{mutante}$, com p_m novas soluções, chamadas soluções mutantes, geradas aleatoriamente tal como na população inicial; e (iii) C_{filhos} , com as $p\text{-}p_e\text{-}p_m$ soluções obtidas a partir do cruzamento (*crossover*) de soluções elite e não elite da geração anterior, g.

As soluções produzidas no processo de cruzamento são chamadas soluções filhas. No BRKGA, o processo de cruzamento ocorre sempre entre um cromossomo (solução) do conjunto elite e um do conjunto não elite. Uma vez que há p_e soluções elite e $p\text{-}p_e$ soluções não elite, onde $p_e < p\text{-}p_e$, a probabilidade de um cromossomo do conjunto elite participar de um cruzamento é superior a probabilidade de um cromossomo do conjunto não elite ser incluído, dada por $1/(p\text{-}p_e)$.

Uma vez selecionados os cromossomos utilizados no cruzamento, é necessário definir como os genes de cada cromossomo serão combinados na solução resultante. O parâmetro de entrada $\rho_e > 0.5$ representa a probabilidade de seleção dos genes do cro-

mossomo pertencente ao conjunto elite. A operacionalização do cruzamento ocorre 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]. Os dois cromossomos incluídos no cruzamento, $c_e \in C_{elite}$ e $c_n \notin C_{elite}$, e o vetor v_a são emparelhados. Quando a i-ésima posição de v_a for menor ou igual a ρ_e, o cromossomo filho, c_f, herda, na posição i, o i-ésimo gene de c_e. Caso contrário, o cromossomo filho herda o i-ésimo gene de c_n. Assumindo exemplo hipotético de um conjunto de dados com 10 objetos e $\rho_e = 0.7$, o processo de cruzamento é representada a seguir:

$$c_e = \{0.31,0.77,0.81,0.49,0.32,0.97,0.72,0.15,0.56,0.92\}$$
$$c_n = \{0.26, \mathbf{0.15}, 0.36,0.41,0.93, \mathbf{0.11}, 0.28, \mathbf{0.56}, \mathbf{0.34}, 0.87\}$$
$$v_a = \{0.58, 0.89, 0.11, 0.41, 0.70, 0.99, 0.43, 0.71, 0.88, 0.58\}$$
$$c_f = \{0.31, 0.15, 0.81, 0.49, 0.32, 0.11, 0.72, 0.56, 0.34, 0.92\}$$

O processo de cruzamento é repetido até que $p\text{-}p_e\text{-}p_m$ cromossomos tenham sido gerados e a população da geração seguinte g+1 esteja completa.

O algoritmo descrito admite o número de grupos como parâmetro de entrada, pois o número de medóides é fixo em cada execução. Porém, no problema de microagregação, o parâmetro de entrada é o nível de agregação k. Considerando a abordagem baseada em grupos de tamanho variável, um intervalo para o número de grupos foi definido e múltiplas execuções do algoritmo foram realizadas. Formalmente, para um conjunto de dados T, com n objetos, e um nível de agregação k, o intervalo para o número de grupos c foi definido como $[\max(2, \lceil n/2k \rceil), \max(2, \lfloor n/k \rfloor)]$. Todos os inteiros compreendidos nesse intervalo foram considerados na obtenção das soluções.

4.3 Medidas de validação

O resultado da aplicação de uma técnica de SDC deve ser avaliado segundo dois aspectos [37], são eles: a perda de informação (do inglês, *information loss* - IL) e o risco de violação do sigilo (do inglês, *disclosure risk* - DR).

Quanto à perda de informação, diversas medidas de avaliação estão disponíveis na literatura, tais como medidas de informação teórica [7], medidas estatísticas [6, 38] e medidas probabilísticas [23]. A fim de possibilitar a comparação com outros métodos, a medida adotada no presente trabalho é a mais utilizada na literatura [4, 5, 11, 22], dada pela Equação (6), apresentada a seguir.

Na abordagem clássica do problema de microagregação, a partição ótima, em termos da soma de quadrados, é aquela que minimiza o valor de SSE (Equação (1)). Uma vez que SST representa a soma de quadrados total, a medida $IL1 = SSE/SST$ representa a perda de informação normalizada, $IL1 \in [0,1]$. Quanto mais próximo de zero, melhor é o processo de partição, no que diz respeito à minimização do SSE.

A medida IL1 está relacionada com o grau de homogeneidade dos grupos obtidos no processo de microagregação. Com o objetivo de quantificar o resultado da transformação do conjunto de dados original T no conjunto protegido T′, ou seja, o impacto da transformação dos valores observados, a medida IL2 [38], descrita na Equação (7), foi também considerada. 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 [8]. Uma medida usada para quantificar o LD [18, 24, 26] é o *Distance-based Linkage Disclosure* (DLD), descrito pela Equação (8).

$$\mathrm{IL1} = \frac{\mathrm{SSE}}{\mathrm{SST}} = \frac{\sum_{i=1}^{c}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^T(x_{ij}-\bar{x}_i)}{\sum_{i=1}^{c}\sum_{j=1}^{n_i}(x_{ij}-\bar{x})^T(x_{ij}-\bar{x})} \qquad (6) \qquad \mathrm{IL2} = \frac{1}{np}\sum_{l=1}^{p}\sum_{j=1}^{n}\frac{\left|x_{jl}-x'_{jl}\right|}{\sqrt{2}S_l} \qquad (7)$$

$$\mathrm{DLD} = \frac{\#\left\{x'_j \mid j=\arg\min_m\left(d\left(x'_j,x_m\right)\right), x'_j\in T'_N, x_m\in T_N\right\}}{n} \qquad (8)$$

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. Esse total é dividido pelo número total de objetos n, resultando em uma proporção, logo DLD $\in [0,1]$.

5 Resultados computacionais

Com o objetivo de avaliar o desempenho do método proposto frente a outros métodos presentes na literatura, um experimento computacional foi conduzido nos três conjuntos de dados comumente utilizados[1]. São eles: Tarragona, Census e EIA, com 834, 1080 e 4092 objetos e 13, 13 e 11 atributos numéricos contínuos, respectivamente. Uma vez que as conclusões foram similares, devido a restrições de espaço, apenas os resultados obtidos para os dois primeiros conjuntos foram apresentados.

A configuração dos parâmetros do BRKGA foi realizada com base no experimento executado por Brito et al. [3]. Os seguintes valores foram definidos para os parâmetros: tamanho da população, p=100; tamanho do conjunto elite, p_e=0.20p; número de soluções mutantes, p_m=0.15p; probabilidade de seleção de genes do cromossomo elite no cruzamento, ρ_e=0.7 e total de gerações, g=100.

Para o nível de agregação k, os seguintes valores foram definidos: k={3, 4, 5, 10, 25, 50, 100}. O intervalo do número de grupos, c, foi definido conforme estabelecido na Seção 4.2, ou seja, $c=[\max(2,\lceil n/2k\rceil), \max(2,\lfloor n/k\rfloor)]$.

O método proposto foi implementado em linguagem R e os resultados produzidos com a sua execução foram comparados com aqueles reportados na literatura para os métodos MDAV [9], CBFS [21], NPN-MHM [4], GSMS [29] e IMHM [25], conforme apresentado na **Tabela 1**[2].

A **Tabela 2** mostra os resultados obtidos pelo método proposto para as medida de perda de informação IL2 e de risco de violação do sigilo DLD combinadas, para ambos os conjuntos. Não foi realizada comparação com os demais métodos pois os respectivos resultados não foram encontrados na literatura.

[1] Foram adotados os conjuntos utilizados em comum pelos métodos considerados.

[2] Os valores de IL1 foram multiplicados por 100, seguindo padrão adotado pelas referências consideradas.

Tabela 1. Comparativo IL1

Conjunto Tarragona							
método	k = 3	k = 4	k = 5	k = 10	k = 25	k = 50	k = 100
MDAV	16.932	19.545	22.461	33.192	46.975	58.526	69.550
CBFS	16.974	-	22.827	33.218	-	-	-
NPN-MHM	17.395	-	27.021	40.183	-	-	-
GSMS	**16.610**	-	21.950	33.230	-	-	-
IMHM	16.931	**19.515**	**21.186**	**30.784**	43.452	56.431	68.194
BRKGA	23.219	25.425	26.427	31.635	**36.676**	**42.126**	**46.667**
Conjunto Census							
método	k = 3	k = 4	k = 5	k = 10	k = 25	k = 50	k = 100
MDAV	5.692	7.494	9.088	14.155	21.402	28.996	39.063
CBFS	5.680	-	8.905	13.896	-	-	-
NPN-MHM	6.350	-	11.344	18.734	-	-	-
GSMS	5.560	-	8.670	13.550	-	-	-
IMHM	**5.367**	**6.858**	**8.417**	**12.228**	**18.613**	25.080	32.678
BRKGA	10.176	11.605	12.469	15.514	19.190	**22.260**	**25.707**

Tabela 2. Medidas IL2 e DLD

Conjunto Tarragona							
medida	k = 3	k = 4	k = 5	k = 10	k = 25	k = 50	k = 100
IL2	6.044	6.625	6.791	7.847	9.460	10.812	12.665
DLD	0.185	0.104	0.102	0.037	0.007	0.007	0.004
Conjunto Census							
medida	k = 3	k = 4	k = 5	k = 10	k = 25	k = 50	k = 100
IL2	8.239	9.799	10.314	12.129	14.294	16.779	18.698
DLD	0.276	0.195	0.144	0.076	0.030	0.016	0.006

6 Conclusão e trabalhos futuros

Em relação à perda de informação medida por IL1, na comparação com outros métodos, o desempenho do método proposto não se mostrou satisfatório para níveis de agregação inferiores a 25. Entretanto, para valores superiores a esse patamar, os resultados foram promissores.

Embora não seja comum adotar valores muito superiores a 10 para o nível de agregação, em aplicações específicas, tais como serviços baseados em localização (do inglês, *Location-Based Services* - LBS), relacionados à comunicação sem fio e tecnologia posicional, o valor de k pode chegar a centenas [30].

Uma observação deve ser feita quanto ao uso da medida IL1 para avaliar o desempenho do método proposto, uma vez que, como descrito na Equação (5), a função objetivo definida considera como protótipo dos grupos o medoide, em vez de o centroide, protótipo considerado na abordagem clássica. Dessa forma, essa medida pode não ser a mais apropriada para esse método.

A medida de perda de informação IL2, que avalia a discrepância entre o conjunto de dados original e o conjunto protegido, resultante do processo de microagregação, apresentou resultados satisfatórios. Para os conjuntos Tarragona e Census, respectivamente, a IL2 máxima (k=n) é 17.64 e 48.89, a perda de informação foi inferior a 45% e 25% para $k \leq 10$. No que diz respeito ao risco de violação de sigilo, os valores

obtidos para a medida DLD foram inferiores a 0.2, com exceção apenas do conjunto Census para k=3. Uma vez que DLD $\in [0,1]$, o método proposto ofereceu excelentes resultados. O risco de violação de sigilo se manteve razoavelmente próximo de zero para todos os níveis de agregação. Entretanto, a análise combinada das duas medidas evidencia o caráter multi-objetivo do problema de microagregação.

Uma desvantagem do método proposto é o tempo de execução, que embora não tenha sido apresentado, foi elevado. Entretanto muitas melhorias podem ser conduzidas a fim de reduzir o custo de processamento, tais como, implementar paralelização, definir metodologia para construir o intervalo do número de grupos de maneira não exaustiva e refinar a configuração dos parâmetros do BRKGA, por exemplo.

A análise dos resultados deixa claro que o problema de microagregação é um problema de otimização multi-objetivo, uma vez que a minimização das perdas de informação pode elevar demasiadamente o risco de violação de sigilo, assim como a minimização do risco pode tornar o conjunto resultante pouco útil, em função de significativa perda de informação. Nesse sentido, há abordagens promissoras na literatura, tal como a proposta por Mortazavi e Jalili [26].

7 Referências

1. Bentley, J. L., Multidimensional binary search trees used for associative searching. Commun. ACM, 18(9), 509-517, 1975.
2. Brand, R., Microdata protection through noise addition. L.N. C. Sci., 2316, 97-116, 2002.
3. Brito, J. and Semaan, G. and Brito, L., Resolução do Problema dos k-medoids via Algoritmo Genético de Chaves Aleatórias Viciadas. Revista Pesquisa Naval, 27, 126-142, 2015.
4. Domingo-Ferrer, J. and Martínez-Ballesté, A. and Mateo-Sanz, J. and Sebé, F., Efficient Multivariate Data-oriented Microaggregation. The VLDB Journal, 15(4), 355-369, 2006.
5. Domingo-Ferrer, J. and Mateo-Sanz, J., Practical Data-Oriented Microaggregation for Statistical Disclosure Control. IEEE Trans. on Know. and Data Eng., 14(1), 189-201, 2002.
6. Domingo-Ferrer, J. and Mateo-Sanz, J. and Torra, V., Comparing SDC Methods for Microdata on the Basis of Information Loss and Disclosure Risk. Pre-proceedings of ETK-NTTS, 2, 807-826, 2001.
7. Domingo-Ferrer, J. and Rebollo-Monedero, D., Measuring Risk and Utility of Anonymized Data Using Information Theory. Proceedings of the EDBT/ICDT, 126-130, 2009.
8. Domingo-Ferrer, J. and Torra, V., A quantitative comparison of disclosure control methods for microdata. Confidentiality, disclosure, and data access: Theory and practical applications for statistical agencies, Elsevier, 2001.
9. Domingo-Ferrer, J. and Torra, V., Ordinal, continuous and heterogeneous K-anonymity through microaggregation. Data Mining and Knowledge Discovery, 11(2), 195-212, 2005.
10. Drechsler, J., Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation. Springer-Verlag, 2011.
11. Fayyoumi, E. and Oommen, B., A survey on statistical disclosure control and microaggregation techniques for secure statistical databases. Software: Practice and Experience, 40(12), 1161-1188, 2010.
12. Glover, F. and Kochenberger, G. (Ed.), Handbook of Metaheuristics. Springer, 2003.
13. Gonçalves, J. and Resende, M., Biased Random-key Genetic Algorithms for Combinatorial Optimization. Journal of Heuristics, 17(5), 487-525, 2011.
14. Hair, J. and Black, W. and Babin, B. and Anderson, R. and Tatham, R., Analise Multivariada de Dados. Bookman, 2009.

15. Han, J. and Kamber, M. and Pei, J., Data mining: concepts and techniques. Morgan Kaufmann Publishers, 2012.
16. Hansen, S. and Mukherjee, S., A Polynomial Algorithm for Optimal Univariate Microaggregation. IEEE Trans. on Know. and Data Eng., 15(4), 1043-1044, 2003.
17. Heaton, W., New Record Ordering Heuristics for Multivariate Microaggregation. Tese (doutorado), Nova Southeastern University, 2011.
18. Herranz, J. and Nin, J. and Solé, M., Kd-trees and the Real Disclosure Risks of Large Statistical Databases. Information Fusion, 13(4), 260-273, 2012.
19. Hundepool, A. and Domingo-Ferrer, J. and Franconi, L. and Giessing, S. and Nordholt, E. and Spicer, K. and Wolf, P., Statistical Disclosure Control. John Wiley & Sons, 2012.
20. Kaufman, L. and Rousseeuw, Peter J., Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley, 1990.
21. Laszlo, M. and Mukherjee, S., Minimum spanning tree partitioning algorithm for microaggregation. IEEE Trans. on Know. and Data Eng., 17(7), 902-911, 2005.
22. Li, Y. and Zhu, S. and Wang, L. and Jajodia, S., A Privacy-Enhanced Microaggregation Method. Proceedings of the Second International Symposium on Foundations of Information and Knowledge Systems, 148-159, 2002.
23. Mateo-Sanz, J. and Domingo-Ferrer, J. and Sebé, F., Probabilistic Information Loss Measures in Confidentiality Protection of Continuous Microdata. Data Mining and Knowledge Discovery, 11(2), 181-193, 2005.
24. Mateo-Sanz, J. and Sebé, F. and Domingo-Ferrer, J., Outlier Protection in Continuous Microdata Masking. Proceedings of the CASC Project Final Conference on Privacy in Statistical Databases, 201-215, 2004.
25. Mortazavi, R. and Jalili, S. and Gohargazi, H., Multivariate Microaggregation by Iterative Optimization. Applied Intelligence, 39(3), 529-544, 2013.
26. Mortazavi, R. and Jalili, S., Preference-based Anonymization of Numerical Datasets by Multi-objective Microaggregation. Information Fusion, 25(C), 85-104, 2015.
27. Mortazavi, R. and Jalili, S., Fast data-oriented microaggregation algorithm for large numerical datasets. Knowledge-Based Systems, 67, 195-205, 2014.
28. Oganian, A. and Domingo-Ferrer, J., On the complexity of optimal microaggregation for statistical disclosure control. Statistical Journal of the UNECE, 18(4), 345-353, 2001.
29. Panagiotakis, C. and Tziritas, G., Successive Group Selection for Microaggregation. IEEE Trans. on Know. and Data Eng., 25(5), 1191-1195, 2013.
30. Rebollo-Monedero, D. and Forné, J. and Soriano, M., An Algorithm for K-anonymous Microaggregation and Clustering Inspired by the Design of Distortion-optimized Quantizers. Data & Knowledge Engineering, 70(10), 892-921, 2011.
31. Semaan, G.S., Algoritmos para o problema de agrupamento automático. Tese (Doutorado). Universidade Federal Fluminense. Niterói, 2013.
32. Solanas, A., Privacy Protection with Genetic Algorithms. Success in Evolutionary Computation, Springer Berlin Heidelberg, 2008.
33. Solé, M. and Muntés-Mulero, V. and Nin, J., Efficient microaggregation techniques for large numerical data volumes. Int. Journal of Information Security, 11(4), 253-267, 2012.
34. Templ, M., Statistical Disclosure Control for Microdata: Methods and Applications in R. Springer International Publishing, 2017.
35. Torra, V. and Navarro-Arribas, G. and Stokes, K., An Overview of the Use of Clustering for Data Privacy. Unsupervised Learning Algorithms, Springer, 2016.
36. UNSTATS, Fundamental Principles of Official Statistics. 2014. Disponível em https://unstats.un.org/unsd/dnss/gp/fundprinciples.aspx.
37. Winkler, W., Examples of easy-to-implement, widely used methods of masking for which analytic properties are not justified. Stat. Research Div., U.S. Bureau of the Census, 2007.
38. Yancey, W. and Creecy, R., Disclosure Risk Assessment in Perturbative Microdata Protection. Inference Control in Statistical Databases, From Theory to Practice, 2002.