PROPOSTA DE INDICADORES DE ESTABILIDADE PARA O ESCALONAMENTO MULTIDIMENSIONAL (MDS)

MOACYR MACHADO CARDOSO JUNIOR
Instituto Tecnológico de Aeronáutica – ITA
Praça Mal. Eduardo Gomes, 50 – São José dos Campos
moacyr@ita.br

RODRIGO ARNALDO SCARPEL
Instituto Tecnológico de Aeronáutica – ITA
Praça Mal. Eduardo Gomes, 50 – São José dos Campos
rodrigo@ita.br

RESUMO

O presente trabalho tem por objetivo demonstrar que o resultado final do Escalonamento Multidimensional (MDS) está sujeito à variabilidade das coordenadas originais e que o ajuste final é dependente do número de objetos, e da variabilidade de suas coordenadas. Foram propostos quatro indicadores de qualidade do ajuste final do MDS submetido a uma perturbação aleatória Normal: Erro médio, Erro médio da origem, índice de Janela esférica de origem – Ijor e índice de Janela definida pelo usuário – Ijmd. Os resultados das simulações confirmam que a qualidade do ajuste é dependente do tamanho da matriz de entrada e da variabilidade das suas coordenadas. Os indicadores propostos são capazes de retratar o efeito da perturbação nas coordenadas do espaço MDS.

PALAVRAS CHAVE. Escalonamento Multidimensional (MDS), Estabilidade, Indicadores. Área de classificação principal (Outras aplicações em PO (OA))

ABSTRACT

This work aims to show that Multidimensional Scaling (MDS) final results are dependent of variability of original coordinates and that the final adjust depends on number of objects e variability of theirs coordinates. Four MDS final adjustment quality indicators were proposed when it was submitted to a Normal perturbation: Mean error, Origin Mean Error, Origin Spherical Window Indicator – Ijor and User defined Window Indicator – Ijmd. The simulation results confirm that adjustment quality is dependent of matrix size and variability of its coordinates. The proposed indicators are able to retract the perturbation effect on MDS space coordinates.

KEYWORDS. Indicators, Multidimensional Scaling (MDS), Stability.

Main area (Other applications in OR (OA)).

1. Introdução

O Escalonamento Multidimensional (MDS) é um método que toma por base a proximidade de objetos, sujeitos ou estímulos e é utilizado para produzir uma representação espacial dos mesmos. (HÄRDLE e SIMAR, 2007). A proximidade expressa a similaridade/dissimilaridade entre objetos. O MDS é uma técnica de redução de dimensão, uma vez que seu objetivo é encontrar um conjunto de pontos em baixa dimensão (usualmente duas dimensões) que reflitam a configuração dos dados em alta dimensão.

Cox e Cox (2001), apresentam uma definição de MDS que chamam de "estreita" parecida com a apresentada acima, e uma "larga" ou seja, MDS pode conter várias técnicas de análise multivariada de dados, e no extremo, ela se refere a qualquer técnica que produza uma representação gráfica dos objetos a partir de dados multivariados.

Borg e Groenen (2005) definem MDS como um método que representa medidas de similaridade (ou dissimilaridade) entre pares de objetos como distâncias entre pontos em um espaço de baixa dimensão. Os autores descrevem quatro propósitos do MDS: a) MDS como um método para representar (dis)similaridade como distâncias em um espaço dimensional com poucas dimensões, afim de tornar os dados acessíveis à inspeção visual e análise exploratória; b) Como uma técnica que permita que se teste de que forma certos critérios pelos quais uma pessoa pode distinguir diferentes objetos de interesse são espelhados em uma correspondente diferença empírica desses objetos; c) Uma aproximação analítica que permita descobrir as dimensões que estão sob os julgamentos de (dis)similaridade; d) Como um modelo psicológico que explica julgamentos de (dis)similaridade em termos de regras que se aproximam de um tipo particular de função de distância.

O MDS baseado usualmente em distâncias Euclidianas calculadas a partir de similaridades é usualmente referido como MDS métrico, e se for baseado somente em similaridades baseadas em julgamentos pessoais, é o MDS não métrico, medido em escala ordinal. (RECHER, 2002).

O MDS métrico e não métrico é utilizado atualmente em diversos campos da ciência e disciplinas, tais como citados por Cox e Cox (2001) na área de biometria, psicologia, ecologia, ergonomia, silvicultura, lexicografia, marketing, turismo e conectividade cerebral. Na Tabela 1, apresenta-se uma pesquisa de utilização da ferramenta MDS nas diversas áreas do conhecimento, no idioma português e inglês, onde se verifica o potencial de uso da mesma.

TABELA 1 - Resultados da pesquisa pelos termos "Escalonamento Multidimensional" ou "Multidimensional Scaling" no Google Scholar, para os anos de 2007-2010

Área	Termo Pesquisado	
	Escalonamento Multidimensional	*"Multidimensional Scaling"*
	N° de trabalhos que fazem uso ou referência ao MDS	
Biologia, Ciências da Vida, Ciências do Ambiente	24	+ de 1000
Negócios, Administração, Finanças e Economia	0	542
Química e Ciências dos Materiais	0	237
Engenharia, Ciências da Computação, Matemática	0	+ de 1000
Medicina, Farmacologia, Ciências Veterinárias	1	610
Física, Astronomia, Ciências Planetárias	1	153
Ciências Sociais, Artes, Humanidades	13	+ de 1000

Fonte: Harzing Publish - (Pesquisa dos autores em 12/04/2010)

O objetivo deste estudo é o de propor indicadores de estabilidade para os resultados do MDS ordinal (não métrico), por meio de um estudo teórico de simulação. O trabalho está estruturado em 6 seções, sendo esta a primeira, a título de introdução, a seção 2 que apresenta o

método MDS métrico e não métrico, a seção 3 uma breve revisão sobre estudos de estabilidade do MDS, seção 4 o método proposto, a seção 5 com os resultados obtidos e finalmente as considerações finais.

2. Escalonamento Multidimensional (MDS)

O MDS métrico pode ser calculado por diferentes métodos. O primeiro método, denominado de clássico, foi introduzido por Torgerson (1952, 1958) e Gower (1966), conforme citado por Wickelmaier (2003) e Borg e Groenen (2005). O MDS clássico também é conhecido como Escalonamento de Torgerson ou ainda Torgerson-Gower (BORG e GROENEN, 2005).

O MDS clássico inicia com uma matriz de distâncias **D** com elementos d_{ij}, onde i, j = 1,....n, e o objetivo é encontrar uma configuração de pontos no espaço p-dimensional, a partir das distâncias entre os pontos, de tal forma que as coordenadas dos n pontos, ao longo da dimensão p, produza uma matriz de distâncias Euclidianas, cujos elementos, estão tão próximos quanto possível aos elementos da matriz de distâncias **D**. No método clássico, a matriz de dissimilaridades é assumida como uma matriz de distâncias.

O algoritmo para solução do MDS métrico envolve álgebra linear e está baseado nos Teoremas de Eckart e Young (1936) e de Young e Householder (1938) citado por Borg e Groenen (2005) e Cox e Cox (2000). O algoritmo está baseado no fato de que a matriz de coordenadas **X** pode ser derivada pela decomposição espectral dos autovalores, a partir da matriz do produto escalar $\mathbf{B}=\mathbf{X}.\mathbf{X}^T$, sendo **B** simétrica e semi-definida positiva. O problema da construção de **B**, a partir da matriz de dissimilaridades **P**, é resolvido pela multiplicação do quadrado das dissimilaridades pela matriz $J = I - \frac{1}{n}11'$, sendo **I** a matriz identidade, n o número de objetos ou estímulos em análise e $11'$ uma matriz (nxn) com 1 em todos os elementos. Este processo é denominado por "Double centering". O método é extremamente popular, uma vez que permite soluções analíticas, não requerendo métodos iterativos. (BORG e GROENEN, 2005).

2.1. MDS não métrico ou ordinal

O objetivo do MDS não métrico, tal qual no métrico é o de encontrar as coordenadas dos pontos em um espaço p-dimensional, de forma que exista uma boa concordância entre as proximidades observadas e as distâncias entre pontos. O desenvolvimento da técnica ocorreu devido a dois pontos fracos do MDS métrico, conforme Fahrmeir e Hamarle, (1984) citado por Hardle e Simar (2007), que são:

a) A definição de uma conexão explicita entre dissimilaridades e distâncias, de forma a permitir calcular as distâncias a partir das dissimilaridades, e

b) A restrição à geometria Euclidiana a fim de determinar a configuração final do objeto.

A idéia do MDS não métrico é estabelecer uma relação menos rígida entre dissimilaridades e distâncias. Dessa forma, suponha que uma função monotônica crescente desconhecida *f*, seja usada para gerar um conjunto de distâncias d_{ij} como uma função das dissimilaridades δ_{ij}, $d_{ij} = f(\delta_{ij})$. Consideramos que *f* tem a propriedade que se $\delta_{ij} < \delta_{rs}$ então $f(\delta_{ij}) < f(\delta_{rs})$. O escalonamento é baseado na ordenação das dissimilaridades. O método não métrico é ordinal.

O MDS ordinal é um caso especial de MDS, e possivelmente o mais importante na prática (COX e COX, 2001). Ele é usado normalmente quando, por exemplo, desejamos obter o julgamento, colocando os objetos em ordem crescente ou decrescente de importância sob a ótica de um avaliador.

A aproximação mais comum usada para determinar os elementos d_{ij} e para obter as coordenadas dos objetos x1, x2, ..., xn no espaço p-dimensional é um processo iterativo, implementado no algoritmo de Shepard-Kruskal.

No primeiro passo do algoritmo, denominado fase inicial, são calculadas as distâncias Euclidianas d_{ij}^0 a partir de uma configuração inicial arbitrária X_0 na dimensão p*, desde que todos os objetos tenham coordenadas diferentes. Uma forma usual de se obter as coordenadas iniciais é pela utilização do MDS métrico. Uma outra forma de obter a configuração inicial é pela geração de n pontos de acordo com o Processo de Poisson na Região $\mathbf{R}^p$, ou seja, simular cada coordenada individual de cada ponto, independentemente a partir de uma distribuição uniforme em [0,1].(COX e COX, 2001).

A configuração obtida deve ser normalizada de forma que seu centróide esteja na origem. Este procedimento é possível uma vez que a função de perda a ser calculada é invariante à translação e expansão uniforme.

O segundo passo, ou fase não métrica, determina as disparidades $\hat{d}_{ij}^0$, a partir das distâncias d_{ij}^0, por meio da construção de uma relação monotônica entre d_{ij}^0 e δ_{ij}, sob a condição de que se $\delta_{ij}<\delta_{rs}$ então $\hat{d}_{ij}^0 \leq \hat{d}_{rs}^0$. Essa condição é chamada de monotonicidade fraca. Para obtenção das disparidades $\hat{d}_{ij}^0$, utiliza-se a aproximação dada pelo algoritmo PAV ("*Pool-adjacent violators*").

Seja $(i_1, j_1) > (i_2, j_2) > ... > (i_k, j_k)$ a ordenação dos valores de dissimilaridade de $k = \frac{n(n-1)}{2}$ pares de objetos. O algoritmo PAV começa com o valor δ_{ij} referente a menor distância d_{ij}^0, o valor adjacente d_{ij}^0 é comparado com cada δ_{ij} para determinar se os mesmos são monotonicamente relacionados aos δ_{ij}´s. Sempre que um bloco consecutivo de valores d_{ij}^0 for encontrado que viole o requerimento da monotonicidade o correspondente valor d_{ij}^0 é recalculado pela média com o valor anterior, de forma a obter um estimador.

Com o objetivo de avaliar o ajuste proposto pelo PAV, o Método de Shepard-Kruskal sugere a avaliação pelo indicador denominado Stress.

O Stress é definido por:

$$Stress = \left(\frac{\sum_{i<j} \left(d_{ij} - \hat{d}_{ij}\right)^2}{\sum_{i<j} d_{ij}^{\,2}} \right)^{\frac{1}{2}} \qquad (1)$$

No terceiro passo, chamada fase métrica, a configuração espacial de X_0 é alterada obtendo-se X_1. A partir de X_1 novas distâncias são calculadas d_{ij}^1 e que estão mais próximas as disparidades $\hat{d}_{ij}^0$ do passo 2.

O objetivo final do algoritmo é o de encontrar as coordenadas que minimizam o Stress e mantem a monotonicidade. O processo é iterativo e as novas coordenadas de cada objeto são obtidas a partir da contribuição de todos os objetos pela seguinte expressão:

$$x_{il}^{NOVO} = x_{il} + \frac{\alpha}{n-1} \sum_{j=1;\, j\neq i}^{n} \left(1 - \frac{\hat{d}_{ij}}{d_{ij}}\right).\left(x_{jl} - x_{il}\right) \qquad (2)$$

com $l = 1,.....p^*$

onde α representa o incremento de cada passo da iteração. Kruskal recomenda começar com α=0,2. O processo iterativo é resolvido por método numérico.

No quarto e último passo, verifica-se se o Stress obtido é suficientemente baixo para parar a iteração, e neste caso a configuração final é obtida.

Resumo dos Passos MDS não métrico

a. Escolha uma configuração inicial;
b. Normalize a configuração;
c. Encontre as distâncias
d. Regressão monotônica para o ajuste das disparidades;
e. Encontrar o gradiente. $\frac{\partial Stress}{\partial x}$, se $\frac{\partial Stress}{\partial x} < \epsilon$, onde ϵ é um valor de parada, este será o valor ótimo.

3. Estabilidade do MDS

Desde o início dos estudos publicados por Kruskal, vários métodos de simulação estudaram o comportamento do Stress, vários deles utilizando simulação de Monte Carlo. (COX e COX, 2001). Estudos de simulação realizados por Stenson e Knoll (1969) e Klahr (1969) *apud* Borg e Groenen, 2005 indicam que o Stress cresce a medida que o número de pontos aumenta, e decresce quando a dimensionalidade do espaço final é aumentada.

Spencer (1979) *apud* Borg e Groenen (2005) aproximou as curvas de variação do Stress em função da dimensão e do número de pontos, conforme a equação:

$$\sigma = 0.001\left(a_0 + a_1 m + a_2 n + a_3 \ln(m) + a_4 \sqrt{\ln(n)}\right)$$

onde: a_0=-524,25; a_1=33,8; a_2=-2,54; a_3=-307,26; a_4=588,35. m= dimensão, variando de 1 a 5e n = número de pontos, variando de 10 a 60.

As oscilações das proximidades podem ser devidas a variações não sistemáticas da atenção, limites de discriminação e processos de ativação e decaimento da memória do julgador, Desta forma Hefner (1958) *apud* Borg e Groenen, (2005), propôs um modelo randômico de forma que as distâncias obtidas não se ajustam em um plano, pois cada distância depende da posição particular do ponto no tempo t, e essas posições não são constantes no tempo. Desta forma, pode-se assumir que existe um erro nos valores de proximidade, e que este erro pode ser modelado aleatoriamente, segundo uma distribuição $N(0,\sigma^2)$.

Para a recuperação da distância com o MDS, considerando o erro, pode-se afirmar que a precisão da recuperação da real distância com o MDS cresce com o número de pontos e decresce com o aumento do erro nos dados e também com o aumento da dimensionalidade do espaço de solução. Este fato é explicado, pois no MDS ordinal (não métrico) as regiões isotônicas decrescem muito rapidamente com o aumento do número de pontos. Quanto maior o erro, pior será a recuperação da distância, pois o erro reduz a correspondência entre as distâncias reais e as proximidades, embora o efeito do erro decresça com o aumento do número de pontos, pois quanto mais pontos ocorrem ou uma sobre estimação ou subestimação do erro gerando um efeito de compensação, e propiciando soluções melhores.

Sibson (1979) apud Cox e Cox (2001) estudou o efeito da perturbação da matriz **B** no resultado dos autovalores e autovetores de **B** e assim na matriz de coordenadas **X**. Assim, para uma perturbação pequena ε, a matriz **B** perturbada pode ser calculada pela equação

$$B(\varepsilon) = B + \varepsilon C + \frac{1}{2}\varepsilon^2 D + O(\varepsilon^3)$$

onde **C** e **D** são matrizes simétricas escolhidas para causar a perturbação. Esta perturbação em **B** causa uma perturbação nos autovalores - λ_i e autovetores - υ_i como segue

$$\lambda_i(\varepsilon) = \lambda_i + \varepsilon\mu_i + \frac{1}{2}\varepsilon^2 u_i + o(\varepsilon^2)$$
$$v_i(\varepsilon) = v_i + \varepsilon f_i + \frac{1}{2}\varepsilon^2 g_i + o(\varepsilon^2)$$

e o autor demonstra que

$$\mu_i = v_i^T C v_i \text{ e } f_i = -(B - \lambda_i I)^T C v_i$$

A perturbação também pode ser aplicada à matriz de distâncias $\mathbf{D}^2$, produzindo a matriz D(ε), na forma

$$D(\varepsilon) = D + \varepsilon F + O(\varepsilon^2)$$

onde F é uma matriz simétrica com elementos da diagonal iguais a zero. Dessa forma, a perturbação induzida em λ_i e v_i será

$$\lambda_i(\varepsilon) = \lambda_i + \varepsilon\mu_i + o(\varepsilon^2)$$
$$v_i(\varepsilon) = v_i + \varepsilon f_i + o(\varepsilon^2)$$

onde $\mu_i = -\frac{1}{2}v_i^T F v_i$ e $f_i = \frac{1}{2}(B - \lambda_i I)^T F v_i + \frac{1}{2}(\lambda_i n)^{-1}(1^T F v_i)1$

A matriz **F** pode ser utilizada para investigar várias perturbações de distâncias $\{d_{rs}\}$. Erros aleatórios podem ser modelados atribuindo uma distribuição de probabilidades qualquer a F estudando os efeitos em μ_i e f_i.

Gower e Krzanowski (1999) *apud* Cox e Cox (2001) propuseram um modelo para estudar agrupamentos dos objetos, e denominaram como análise de distâncias para investigar a estrutura intra grupos e entre grupos, de forma semelhante à ANOVA. Desta forma, as dissimilaridades, tratadas como distâncias, são desdobradas em soma dos quadrados intra agrupamento e soma dos quadrados entre agrupamentos. (T = W + B). Entretanto diferentemente da ANOVA, não existe uma teoria de distribuição de probabilidades estabelecida para a análise de distâncias.

Krzanowski (2006), afirma que muitas análises de MDS findam com a produção de uma representação pictórica em duas ou três dimensões, mas na verdade este é apenas o ponto de partida e assim propôs um método para análise de sensibilidade do MDS métrico que consiste em realizar n+1 análises de coordenadas principais (PCA), uma vez na matriz completa e uma vez para cada omissão de uma linha/coluna desta matriz. As coordenadas dos pontos são calculadas pela decomposição espectral. Cada versão reduzida da matriz é aumentada calculando-se as coordenadas do ponto omitido, usando uma fórmula que requer somente a multiplicação de matrizes já calculadas. O autor sugere novos estudos para avaliar qual modelo probabilístico poderia ser utilizado para inferência, e sugere que estes estudos sejam realizados com bootstrapping.

4. Método para avaliar estabilidade

O método aplicado neste trabalho pode ser dividido em duas etapas: Verificação do desempenho do MDS, em função do número de objetos, e na presença de perturbação, e Simulação da estabilidade dos resultados obtidos com a matriz original e com a introdução de uma perturbação em cada objeto individualmente.

Para verificar o desempenho do MDS Ordinal (não métrico) em relação ao número de objetos(N) e também avaliar o desempenho em relação à variabilidade das coordenadas dos objetos gerou-se matrizes de pontos com N assumindo os valores de 10, 50 e 200. Cada matriz **D** gerada era composta por N linhas (n° de objetos) e 10 colunas de atributos. A distribuição das coordenadas seguiu uma distribuição uniforme com d_{ij} [0,10].

A Matriz **D** obtida foi então submetida ao Escalonamento Multidimensional, utilizando o procedimento KYST (Kruskal-Young-Shepard-Togerson) para escalonamento não métrico. A

dimensão final variou desde 1 até 9, obtendo-se o valor do Stress final. Posteriormente, nesta mesma Matriz **D**, foi introduzida uma perturbação das coordenadas com valores aleatórios gerados segundo uma distribuição Normal com média zero e variância $\sigma^2 \sim N(0,\sigma^2)$. Os valores de σ utilizados foram 1, 3 e 10.

Os valores obtidos para 500 simulações foram submetidos à análise de variância (ANOVA) para testar a hipótese nula.

$$HO: \mu_{dfm1} = \mu_{dfm2} = \mu_{dfm3} = \cdots = \mu_{dfm9}$$

ou seja, dada uma perturbação nas coordenadas originais da matriz **D** a hipótese nula é de que os valores de Stress obtidos a partir da simulação não produziriam diferenças estatísticas entre as dimensões consideradas. Considerou-se neste caso α de 10% de significância.

Os valores médios de Stress obtidos foram ainda comparados com os valores obtidos para a matriz **D** original, utilizando teste "t", assumindo que a distribuição de Stress é Normal.

4.1. Estabilidade

Para verificação da estabilidade foram geradas duas matrizes com N=10 e N=50, também com 10 colunas de atributos com valores $d_{ij} \in [0,10]$. Para estas matrizes foram então obtidas a configuração em baixa dimensão p=2, uma vez que se trata da dimensão que permitem a visualização dos dados e de maior utilização.

Com a configuração final, simulou-se então uma perturbação das coordenadas originais, segundo uma distribuição Normal $\sim N(0,\sigma^2)$, dos objetos individualmente, escolhe-se um objeto e provoca uma perturbação nas coordenadas iniciais, com a finalidade de comparar a distribuição do erro entre a coordenada do espaço MDS final e as novas coordenadas obtidas por simulação. Para mensurar o efeito da variabilidade são propostos alguns indicadores:

1 - O erro médio (ERM), que representa a distância média Euclidiana das coordenadas do objeto Oi em relação às coordenadas obtidas para o conjunto de 100 pontos simulados, o que seria considerado uma medida de desempenho do MDS. Quanto mais próximo a zero, melhor;

2 – O erro médio em relação a origem (ponto com coordenadas 0,0) – (ERM(O)), que representa a distância média Euclidiana das coordenadas simuladas em relação à origem. Quanto mais próximo este valor for da distância Euclidiana do objeto Oi, melhor será o resultado.

3 – Índice de Janela Esférica Origem– (Ijor), este valor representa o percentual de pontos simulados que recaem sobre a mesma região do espaço do ponto original. Para o caso do p=2, equivale ao número de pontos que ficaram em uma Janela circular com raio igual a distância Euclidiana do objeto sob análise e o centro das coordenadas. Tal indicador permite identificar as chances do objeto permanecer no espaço contido entre a origem e o ponto sem perturbação; Quanto mais próximo de 100, melhor o valor.

4 – Índice de Janela definida pelo usuário – (IJmd), refere-se a distância pré-determinada pelo usuário, para avaliar as chances do ponto permanecer em uma dada região do espaço final do MDS. Também quanto mais próximo de 100, melhor a estabilidade.

Na Figura 1, pode visualizar os indicadores 3 e 4 propostos.

5. Resultados

Os resultados obtidos para o Stress nas nove dimensões estão sumarizados na Figura 2, para os dados de Stress obtidos a partir da matriz **D** original (Figura 2-A) e para Stress obtidos com perturbações σ de 1, 3 e 10 (Figura 2-B, 2-C e 2-D) . Pode-se visualizar pela Figura que no caso da Figura 2-A, existe claramente a formação de "cotovelo", indicando a dimensão 5 como a ideal para representar os dados, o mesmo acontecendo para as demais figuras, resultando que a perturbação não altera esta interpretação.

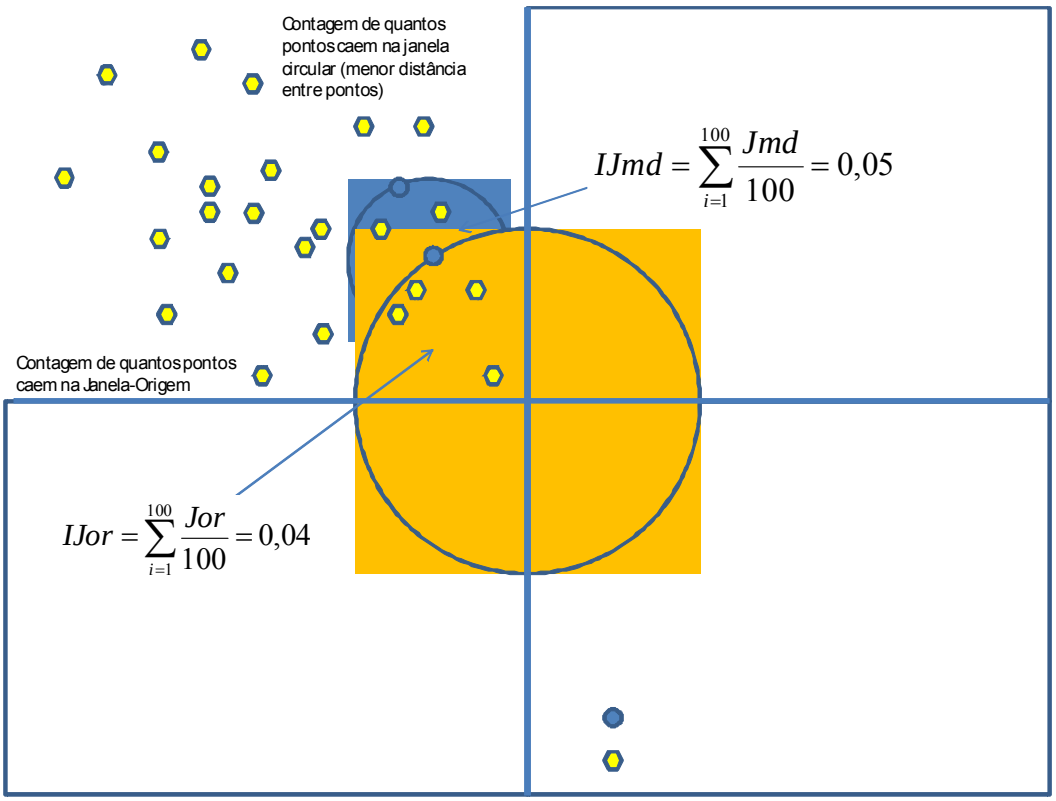

FIGURA 1 - Representação dos Indicadores de Janela de Origem e Janela definida pelo usuário

Na Tabela 2 estão os resultados da ANOVA para Stress em função da dimensão e perturbação. Os resultados obtidos confirmam que o Stress obtido em cada dimensão difere estatisticamente das outras dimensões, ou seja, a escolha da dimensionalidade para representar os dados em situações de variabilidade das coordenadas é estatisticamente significativa com α=10%.

TABELA 2 - Resultados da ANOVA para os valores de Stress obtidos após perturbação da matriz de coordenadas originais, com DP=1, 3 e 10, para N=10, 50 e 200

N - DP	H0	P(>F)
10 – DP=1	$HO: \mu_{10}^{DP=1}_{DIM1} = \cdots = \mu_{10}^{DP=1}_{DIM9}$	2.2e-16 - Não rejeita H0
10 – DP=3	$HO: \mu_{10}^{DP=3}_{DIM1} = \cdots = \mu_{10}^{DP=3}_{DIM9}$	2.2e-16 - Não rejeita H0
10 – DP=10	$HO: \mu_{10}^{DP=10}_{DIM1} = \cdots = \mu_{10}^{DP=10}_{DIM9}$	2.2e-16 - Não rejeita H0
50 – DP=1	$HO: \mu_{50}^{DP=1}_{DIM1} = \cdots = \mu_{50}^{DP=1}_{DIM9}$	2.2e-16 - Não rejeita H0
50 – DP=3	$HO: \mu_{50}^{DP=3}_{DIM1} = \cdots = \mu_{50}^{DP=3}_{DIM9}$	2.2e-16 - Não rejeita H0
50 – DP=10	$HO: \mu_{50}^{DP=10}_{DIM1} = \cdots = \mu_{50}^{DP=10}_{DIM9}$	2.2e-16 - Não rejeita H0
200 – DP=1	$HO: \mu_{200}^{DP=1}_{DIM1} = \cdots = \mu_{200}^{DP=1}_{DIM9}$	2.2e-16 - Não rejeita H0
200 – DP=3	$HO: \mu_{200}^{DP=3}_{DIM1} = \cdots = \mu_{200}^{DP=3}_{DIM9}$	2.2e-16 - Não rejeita H0
200 – DP=10	$HO: \mu_{200}^{DP=10}_{DIM1} = \cdots = \mu_{200}^{DP=10}_{DIM9}$	2.2e-16 - Não rejeita H0

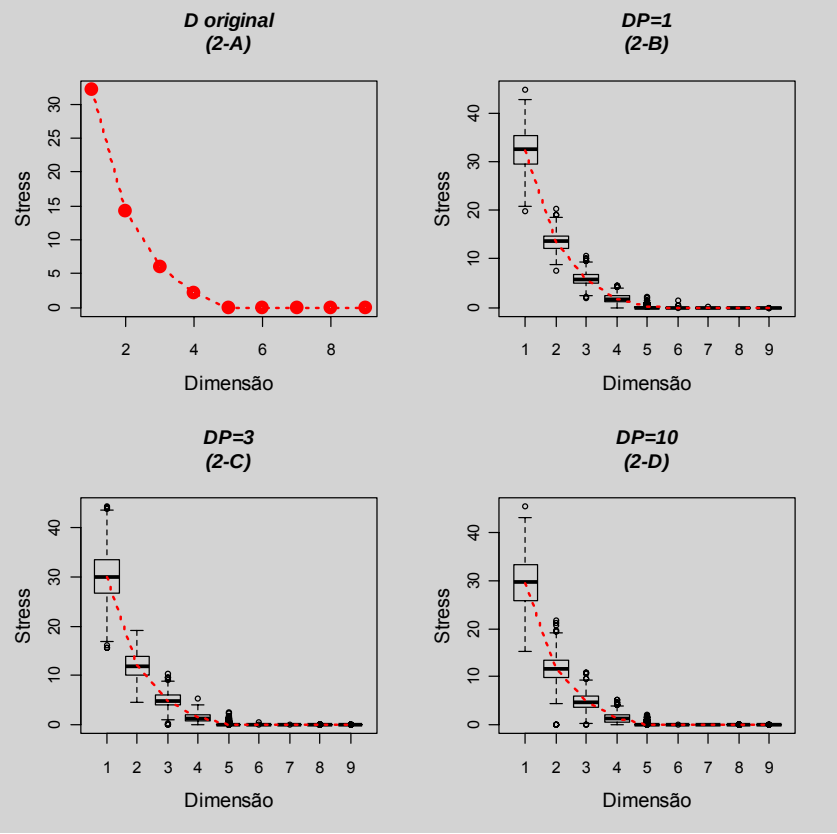

FIGURA 2 - Valores do Stress em função da dimensão da configuração final do MDS, para coordenadas originais e para coordenadas com perturbação DP=1, 3 e 10, N=10

Para a comparação do valor de Stress obtido para a matriz de distância sem perturbação, e com perturbação o teste "t" indicou que somente para o caso unidimensional e DP=1 não se pode afirmar que o Stress é diferente ao nível de significância de 5%. Os demais casos todos são estatisticamente diferentes dos valores com perturbação, onde se conclui que a variabilidade dos dados de origem produz efetivamente valores menores de Stress, conforme Tabela 3.

TABELA 3 - Teste "t" para médias de Stress obtidos na simulação e valor original, para 9 dimensões com perturbações de 1, 3 e 10, para N=10

Dimensão	Valor Original Stress	DP=1 Stress	DP=3 Stress	DP=10 Stress
1	32,1350	**31.83 -32.61***	29.63 - 30.55	29.13 - 30.07
2	14,2524	13.38 - 13.71	11.83 - 12.30	11.47 - 11.99
3	6,0372	5.69 - 5.92	4.79 - 5.08	4.69 - 5.00
4	2,1391	1.75 - 1.91	1.38 - 1.54	1.35 - 1.53
5	0,0014	0.0918 - 0.1379	0.0818 - 0.1342	0.1059 - 0.1596
6	0.0087	0.0025 - 0.0153	0.0027 - 0.0069	0.0038 - 0.0051
7	0,0066	0.0032 -0.0042	0.0026 - 0.0033	0.0028 - 0.0034
8	0,0005	0.0019 - 0.0024	0.0012 - 0.0017	0.0011 - 0.0016
9	0,0000	-	-	-

*no Intervalo confiança 95%

Na Tabela 4 e Figura 3 apresentam-se os resultados da simulação do MDS, com número de objetos N=10 e perturbação com desvio padrão de 0,05 (desvio muito pequeno). Neste caso utilizou-se a Janela esférica IJus = 1.

TABELA 4 - Resultados da simulação do MDS, com N=10, DP=0,05 e DP=1

	N=10, DP=0,05				N=10, DP=1			
Objeto	**ERM**	**ERM(0)**	**Ijmd**	**Ijor**	**ERM**	**ERM(0)**	**Ijmd**	**Ijor**
1	1.118448	4.481175	62	5	13.04913	15.2855	0	2
2	2.091351	6.096864	8	6	14.24569	17.78546	0	19
3	0.728001	2.609946	91	0	16.16075	18.08866	0	0
4	2.481348	8.284462	0	0	10.79399	14.32541	0	44
5	1.634751	9.913713	26	0	10.30236	17.63251	0	19
6	0.829886	3.361296	72	3	9.455958	11.54007	0	7
7	0.572214	2.232586	100	37	9.093804	10.87729	1	6
8	1.919112	7.435356	0	96	15.82179	19.28568	0	45
9	1.435376	4.416793	22	99	18.09389	19.67344	0	21
10	2.90112	7.713136	0	90	9.911963	15.19052	0	28

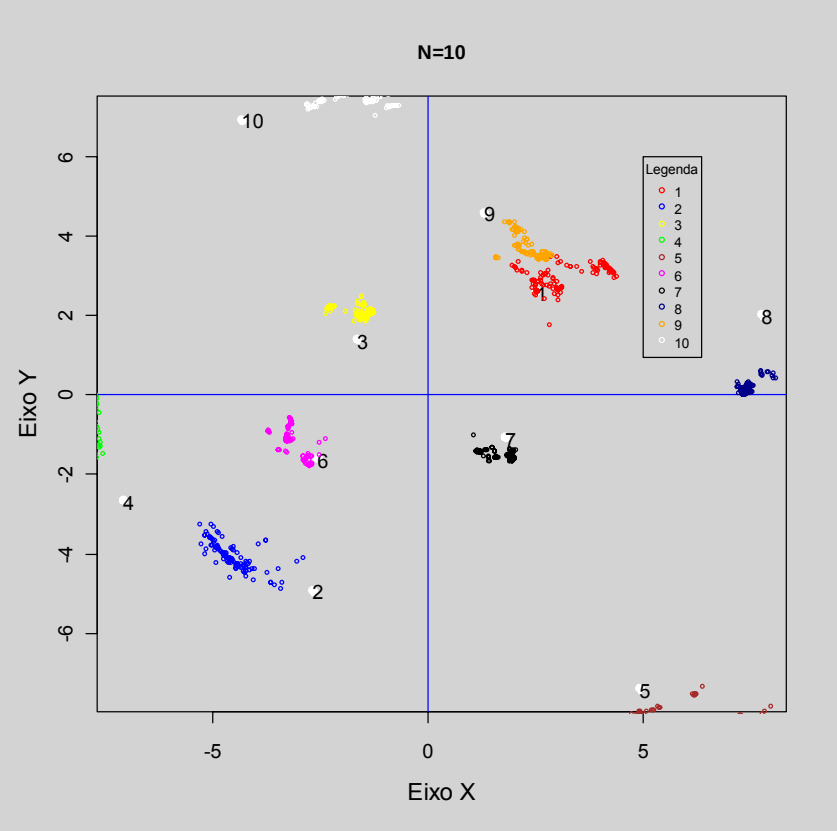

FIGURA 3 - Representação dos 10 pontos no espaço bidimensional, e os respectivos pontos obtidos com a simulação da perturbação nas coordenadas originais, com DP=0.05

Verifica-se que apesar da pequena perturbação provocada nas coordenadas originais, a dispersão dos pontos pode ser significativa, como por exemplo, é o caso dos objetos n° 5, 6, 8, 9 e 10.

Para o objeto n° 10, por exemplo, obteve-se um IJmd = 0, o que significa que 0% dos pontos simulados recaíram no espaço circunscrito ao circulo de raio = 1 em torno do objeto 10. Para o IJor, obteve-se que 90% dos pontos ficaram no espaço delimitado pelos eixos X e Y e um circulo de raio igual a distância entre a origem e o objeto n° 10, no caso de 8,18 unidades de distância. Os indicadores ERM = 2,90 e ERM(O)=7,71, indicam respectivamente que a distância

média do objeto original e os demais 100 pontos simulados é de 2,90 unidades de distância e que a distância média dos 100 pontos à origem é de 7,71, próxima a distância real de 8,18

O objeto n° 1 ficou com Ijmd de 62%, ou seja, 62% de chances dos pontos recaírem na janela de distância de raio igual a 1, porém em relação a origem somente 5% ficaram naquele espaço. O ERM reflete este resultado.

No caso da simulação com N=10 e DP=1, obteve-se os dados da Tabela 5 e Figura 4. Neste caso verifica-se que a instabilidade da configuração final do MDS é maior ainda, conforme pode ser observado para o mesmo objeto n° 10, que continua com Ijmd=0, ou seja, 0% de chances de 1 ponto qualquer da simulação cair no espaço circunscrito na esfera com raio 1. E apenas 28% dos pontos ficam entre a origem, e a esfera com raio igual a distância ao objeto original n°10. Os valores de ERM e ERM(0) refletem esta tendência. Se estivermos interessados no objeto n° 7, nada se pode concluir sobre sua proximidade à origem, uma vez que o ERM(0) foi muito superior a distância real (10,87 e 3,03). A mesma conclusão serve para os demais pontos. O ponto 4 e 8, embora apresentem respectivamente Ijor de 44% e 45%, o que significa que quase metade dos pontos estão situados na região entre a origem e o objeto original, porém com um ERM alto, indicando alta variabilidade.

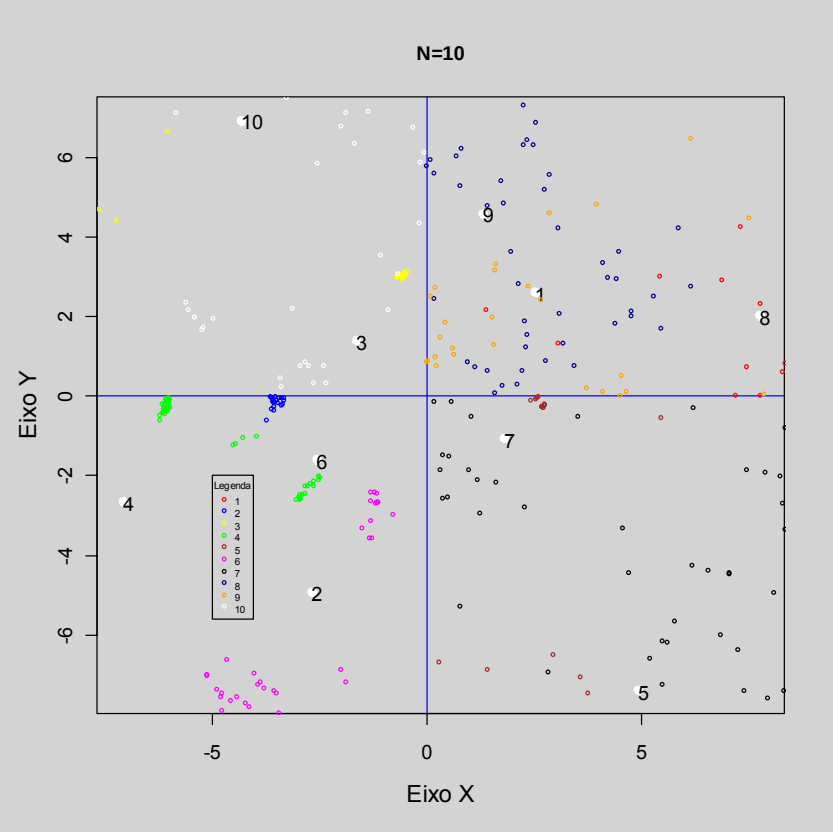

FIGURA 4 - Representação dos 10 pontos no espaço bidimensional, e os respectivos pontos obtidos com a simulação da perturbação nas coordenadas originais, com DP=1

Finalmente testou-se a configuração final dos 50 pontos com desvio padrão de 0,05 e 1, onde foram obtidos os valores da tabela 5.

TABELA 5 - Resultados da simulação do MDS, com N=50, DP=0.05 e DP=1

	N=50, DP=0.05				**N=50, DP=1**			
Objeto	**ERM**	**ERM(0)**	**Ijmd**	**Ijor**	**ERM**	**ERM(0)**	**Ijmd**	**Ijor**
1	0.596798	8.778375	89	65	51.7562	59.88463	0	5
2	0.595995	3.296935	91	10	36.5394	38.84794	0	33

3	0.618873	5.92151	100	2	11.41551	11.5011	0	86
4	0.723121	4.978059	93	95	24.51407	27.26094	0	46
5	0.746665	1.892902	83	25	22.89064	23.62468	0	0
6	1.049423	4.452513	46	93	24.80668	27.46718	0	23
7	0.198861	6.918278	100	32	9.156561	11.09937	0	36
8	1.896242	7.338119	1	0	12.51922	15.86657	0	18
9	0.901551	4.004377	80	93	9.255437	12.10797	1	38
10	0.29594	2.324212	100	2	14.94298	16.53664	7	10

Obs: representados apenas os 10 primeiros pontos.

Verifica-se que para o caso de N=50, as regiões isotônicas diminuem drasticamente, de forma que os pontos, apesar de submetidos aos mesmos desvios anteriores tem muito menos regiões do espaço para serem alocados, resultando em valores de Stress maiores, e, por conseguinte este fato se reflete nos indicadores de dispersão ERM e no Ijmd e Ijor. Por exemplo, o ponto 4, praticamente 93 % dos pontos simulados ficam em uma janela com distância de 1, e 95% ficam na janela circular descrita pela distância do ponto 4 à origem, que é de 5,20. O ERM(O) é de 4,97, que fica muito próximo da distância à origem e ERM = 0,72, o que reflete a pequena dispersão em torno do ponto. Já para o DP=1, a janela Ijmd, com tamanho igual a 1 não tem nenhum ponto, o que é refletido no alto valor do ERM, e da mesma forma o Ijor = 46, e ERM(O)=27,2 denotando que apenas 46% dos pontos permanecem na região descrita pela janela circular de raio igual a 5,20, refletindo também alta variabilidade pelo ERM(O).

6. Conclusões

Os resultados obtidos neste estudo permitem afirmar que a variabilidade das coordenadas dos objetos originais em alta dimensão interferem no resultado final do MDS não métrico, produzindo valores menores de Stress, exceto quando a configuração final é obtida em apenas 1 dimensão, o que pode ser explicado por problemas de ótimos locais do caso unidimensional. O ajuste obtido em baixa dimensão produz grandes dispersões em termos do Stress obtido, o que reflete no objetivo do MDS que é produzir distâncias no espaço de baixa dimensão tão próximas quanto possível das do espaço de alta dimensionalidade.

Quanto aos indicadores propostos, ERM, ERM(O), Ijor e Ijmd, os mesmos se mostram promissores para avaliar a estabilidade do resultado do MDS, sob condições de variabilidade das coordenadas originais refletindo a disposição final obtida.

Como continuidade deste trabalho, no entanto, os indicadores propostos devem ser testados com dados reais, e assumindo a variabilidade inerente ao conjunto de dados.

7. Referências

Borg, I.,Groenen, P. J. *Modern Mutidimensional Scaling: Theory and Applications*. New York: Springer, 2005

Cox, T. F., Cox, M. A. *Multidimensional Scaling*, 2º Ed. (2ª Ed. ed.). London: Chapman & Hall/CRC, 2000

Hardle, W., Simar, L. *Applied Multivariate Statistical Analysis*, 2ª Ed. Berlin: Springer, 2007

Harzing, A.W. (2010) Publish or Perish, V.(2.8.3644), Disponível em: www.harzing.com/pop.htm

Krzanovski, W. J. (2006). Sensitivity in metric Scaling and Analysis of Distance. *Biometrics* , 239-244.

Rencher, A. *Methods of Multivariate Analysis*. New York: John Wiley & Sons, 2ª Ed, 2002