# Incorporando a variabilidade no processo de identificação do modelo de máximo global no *Grade of Membership* (GoM): considerações metodológicas

**Gilvan Ramalho Guedes***
**Pamila Cristina Lima Siviero****
**André Junqueira Caetano*****
**Carla Jorge Machado******
**Eduardo Brondízio*******

*A disponibilidade de bases de dados cada vez mais complexas e multidimensionais é um dos principais motivadores para o aumento do número de estudos que utilizam análises multivariadas baseadas em lógica de conjuntos nebulosos. Apesar da disseminação do método Grade of Membership nos trabalhos empíricos brasileiros da área de ciências sociais e saúde, questões relativas à identificabilidade e estabilidade dos parâmetros finais estimados pelo programa GoM 3.4 não foram suficientemente aprofundadas. Dada a relevância de se obterem parâmetros únicos e estáveis, Guedes et al. (2010) propuseram um procedimento empírico para localizar um modelo de máximo global (MG) com parâmetros estáveis. Entretanto, seu localizador de MG não incorpora qualquer medida de variabilidade. Neste artigo, tal limitação é contornada por meio da utilização de uma estatística de ponderação – Máximo Global Ponderado (MGP) – semelhante ao coeficiente de variação. Esse indicador busca não penalizar de forma desproporcional situações nas quais os desvios médios, apesar de diferentes de zero, são muito pequenos. Apresentam-se evidências de que o localizador MGP reduz a distância do modelo identificado à real estrutura latente dos dados em análise, quando comparados ao modelo identificado pelo localizador não ponderado, MG.*

**Palavras-chave:** Grade of Membership. Máximo Global Ponderado. Variabilidade. Identificabilidade.

* Doutor em Demografia, professor adjunto da Pós-Graduação em Gestão Integrada do Território/Univale, cientista colaborador do Environmental Change Initiative/Brown University; cientista colaborador do Anthropological Center for Training on Global Environmental Change/Indiana University.
** Mestre e doutoranda em Demografia do Centro de Desenvolvimento e Planejamento Regional – Cedeplar/UFMG.
*** Ph.D in Sociology, professor adjunto nível III da Pontifícia Universidade Católica de Minas Gerais – PUC/Minas, pesquisador associado do Centro de Desenvolvimento e Planejamento Regional – Cedeplar/UFMG.
**** Ph.D in Population Dynamics, professora adjunta nível III do Centro de Desenvolvimento e Planejamento Regional – Cedeplar/UFMG.
***** Ph.D in Anthropology, professor e chefe do Department of Anthropology/Indiana University. Codiretor do Anthropological Center for Training and Research on Global Environmental Change/Indiana University Research Scholar – Cipec.

## Introdução

Nos últimos anos, cresceu rapidamente a demanda por algoritmos capazes de encontrar estruturas implícitas aos dados, em resposta à disponibilidade de bancos de dados mais complexos e multidimensionais (VELOSO et al., 2001).

Apesar da existência de diversas técnicas multivariadas (tais como algoritmos que assumem pertencimento exato de indivíduos aos conjuntos, como K-Means, Análise Fatorial e Componentes Principais, e algoritmos que assumem pertencimento múltiplo a conjuntos nebulosos, como FANNY e Fuzzy K-Means), poucas são as que explicitamente fornecem parâmetros para a heterogeneidade amostral no nível das associações entre as categorias das variáveis analisadas. O método *Grade of Membership* (GoM) supre essa lacuna ao estimar a heterogeneidade individual com base em graus de pertencimento a perfis de referência que emergem da estrutura implícita aos dados (MANTON et al., 1994). O GoM, portanto, permite que esse parâmetro individual represente as partições observadas de variáveis latentes (CAETANO; MACHADO, 2009).

Diante da complexidade das bases de dados mais recentes, especialmente as que envolvem dados quantitativos e qualitativos ao longo do tempo (longitudinais) e em diversas escalas de análise (multiníveis), a parametrização da heterogeneidade implícita por meio de partições contínuas permite ao pesquisador evitar a arbitrariedade das categorizações tradicionais e dos grupamentos estanques. Nesse sentido, os algoritmos baseados em lógica nebulosa (*fuzzy sets*), como FANNY e GoM, são mais desejáveis (GILES, 1988).

A principal diferença quantitativa entre algoritmos FANNY, Fuzzy K-Means e GoM ocorre em função de o primeiro ser utilizado para variáveis contínuas (por exemplo, renda e gasto), ao passo que o GoM usa variáveis discretas (como classes de renda e de gasto). Apesar de algoritmos tais como FANNY parecerem mais atraentes por não incorrerem na perda de variabilidade com a categorização de variáveis contínuas (KAUFMAN; ROUSSEEUW, 1990), seu fator final não utiliza a associação no nível das categorias de resposta, como o GoM, representando perda da variabilidade presente na estrutura latente dos dados (MANTON et al., 1994; GUEDES et al., 2009a). Os algoritmos Fuzzy K-Means e FANNY apenas criam um parâmetro adicional para o indivíduo, mas não permitem que sejam obtidas probabilidades associadas às categorias de resposta.

Apesar da sua vasta aplicabilidade em estudos no Brasil (por exemplo, SAWYER et al., 2002; DRUMOND et al., 2007; MELO, 2007; ALVES et al., 2008; GUIMARÃES et al., 2009), foi só recentemente que questões relativas à identificabilidade e estabilidade final dos parâmetros estimados pelo programa GoM 3.4 foram levantadas e suas soluções propostas. A seção seguinte dedica-se a uma breve revisão dos antecedentes empíricos e dos avanços metodológicos recentes propostos no Brasil.

## Antecedentes metodológicos

Muitos trabalhos empíricos, especialmente voltados para a área da saúde (SAWYER et al., 2002; ALVES et al., 2008) e mercado de trabalho (MELO, 2007), têm utilizado a ferramenta como estratégia empírica para identificação de perfis.

Mais recentemente, a aplicação do GoM tem incorporado áreas como hierarquias urbanas (GARCIA et al., 2007; GUEDES et al., 2009a, 2009b), pobreza (GUEDES et al., 2009c), migração e meio ambiente (GUEDES, 2010; SANTOS, 2010), além do importante avanço na sua interlocução com abordagens qualitativas (MIRANDA-RIBEIRO et al., 2007). Essa é uma grande promessa para os tratamentos multimétodo (PEARCE, 2002), especialmente no recrutamento de participantes em grupos focais, tendo como ponto de partida perfis multidimensionais (MIRANDA-RIBEIRO et al., 2007), e na utilização de dados qualitativos para o fornecimento de matrizes iniciais de probabilidade de pertencimento aos perfis multidimensionais de referência (GUEDES, 2010).

A despeito da larga utilização empírica do GoM nas ciências sociais brasileiras, especialmente entre os demógrafos e estudiosos da área de saúde, questões

relativas à identificabilidade e estabilidade dos parâmetros estimados foram até recentemente negligenciadas. Em suma, a ideia da identificabilidade é a de que um modelo deve convergir para uma solução única; caso contrário, não pode ser considerado "confiável".

Na literatura internacional, bem como na nacional, não existem trabalhos voltados para o desenvolvimento de indicadores que auxiliem a busca sistemática de um modelo "confiável". Em recente nota metodológica, Caetano e Machado (2009) introduziram um conceito teoricamente relevante: a questão de identificabilidade, que se refere à capacidade de um método gerar parâmetros solucionáveis e únicos (GILES, 1988). Os autores argumentam que, devido à dependência de uma matriz inicial de probabilidades, o processo iterativo utilizado pelo algoritmo GoM (WOODBURRY; CLIVE, 1974) não é capaz de, por si só, criar uma solução única para os dois parâmetros estimados: $\lambda_{kjl}$ (probabilidade de pertencimento da categoria *l* da variável *j* ao perfil extremo *k*) e $g_{ik}$ (grau de pertencimento do indivíduo *i* ao perfil extremo *k*).

Com base neste trabalho, Guedes et al. (2010) desenvolveram um método empírico de localização de um modelo final identificável, ou seja, com solução única para seus parâmetros. Os autores propõem uma medida chamada $DM_{kjl,r}$ (Estatística de Desvio em Relação à Média) para os parâmetros $\lambda_{kjl}$. Como se está em busca de um modelo identificado, no qual os parâmetros variem muito pouco, essa medida indica a variação do $\lambda_{kjl}$ entre uma execução e outra – r e (r+1). Nesse sentido, o desejável é que se obtenha uma quantidade elevada de números de DM iguais a zero, indicando que grandes diferenças já não são observadas. No mesmo trabalho, os autores identificam problemas relativos à estabilidade de $\lambda_{kjl}$ e $g_{ik}$ resultantes da incapacidade do processo de convergência de encontrar o valor máximo da função de verossimilhança para qualquer estrutura final (qualquer modelo de ordem K).

Embora o procedimento de estabilização dos parâmetros tenha critério único e preciso, o critério de localização do modelo de máximo global sugerido pelos autores é baseado no ordenamento de execuções aleatórias (r) com número decrescente de $DM_{kjl,r} = 0$ ao longo das *L* categorias relativas às *J* variáveis no perfil extremo *k*. Como a ordem (posição) da execução com o maior número de desvios médios nulos varia por perfil extremo, a identificação do máximo global é baseada na média da posição por perfil, uma vez que é necessário selecionar a matriz de probabilidades de uma única execução (rodada). Apesar de ser um critério relativamente simples, a solução encontrada desconsidera qualquer medida de variabilidade na identificação do modelo ótimo. Ou seja, em um caso extremo, se houver desvios médios muito pequenos (variabilidade geral pequena), mas nenhum igual a zero, esse conjunto de desvios será penalizado em decorrência da contagem de desvios iguais a zero (contagem nula). Por outro lado, se houver um conjunto de desvios com alguns muito elevados, mas também com alguns iguais a zero, a contagem de desvios nulos será maior do que na situação anterior e esse conjunto de desvios será menos penalizado. Com base nesta motivação propõe-se, neste trabalho, uma extensão da identificabilidade empírica, incorporando a incerteza (variabilidade) sobre a localização dos parâmetros finais.

## Incorporando variabilidade à medida de identificação do máximo global

Um dos pontos sensíveis ao procedimento de identificabilidade sugerido por Guedes et al. (2010) refere-se à inexistência de um critério que penalize o posicionamento de cada execução aleatória, *r*, com alguma medida de variabilidade. Nesse sentido, a localização do modelo de máximo global proposto em trabalho anterior dá pesos iguais àqueles perfis com o mesmo número de desvios médios iguais a zero, *independentemente* da variabilidade desse DM em cada execução.

Conforme argumentado pelos autores, o valor de DM varia em função de três fatores: número de execuções (*R*); relação $(\lambda_{kjl} - \lambda_{kjl(médio)})$; e número de categorias (*L*). A estatística de identificabilidade proposta

por Guedes et al. (2010), portanto, tem propriedades assintóticas claras, com sua variabilidade decrescendo com o aumento de $R$ e $L$. O procedimento de identificabilidade, por seu turno, ao ser baseado no posicionamento do $\sum \mathrm{DM}_{kjl,r} = 0$ ao longo das categorias $L$ para cada $k$, também depende assintoticamente do número de categorias das variáveis internas ao modelo final.

A dependência assintótica da localização do máximo global em relação ao número de categorias reforça um dos pressupostos do método GoM, de que o aumento de variáveis e, consequentemente, de categorias contribui para um delineamento mais preciso dos perfis extremos (os quais dependem dos valores finais de $\lambda_{kjl}$) (MANTON et al., 1994). Dada a relevância de se chegar ao modelo que descreve mais fidedignamente a estrutura implícita aos dados (e, portanto, encontrar os $\lambda_{kjl}$ mais próximos do máximo global), propõe-se, neste artigo, um procedimento revisado do máximo global (MG) empírico sugerido por Guedes et al. (2010), por intermédio de um localizador de *Máximo Global Ponderado (MGP)*.

O estimador MGP utiliza como medida de variabilidade o desvio padrão dos DM ao longo das $R$ execuções, por perfil extremo $k$. Partindo das estimativas tradicionais dos lambdas ($\lambda_{kjl}$) para cada perfil extremo $k$, basta calcular a média desses lambdas ao longo de $R$ execuções para cada uma das $L$ categorias. A Tabela 1 representa, em forma matricial, a organização dos lambdas e da média necessária para o cálculo da estatística de Desvio em Relação à Média (DM) proposta por Guedes et al. (2010).

A estatística DM é facilmente calculável com base na média dos lambdas por categoria em cada um dos $k$ perfis extremos, bastando subtrair cada um dos valores médios em relação ao valor estimado de $\lambda_{ljk}$ para cada uma das $R$ execuções de uma mesma categoria, $l$. A fórmula de DM está disponível no artigo de Guedes et al. (2010). Uma vez obtidos os valores de DM, soma-se, ao longo das $L$ categorias das $J$ variáveis, o número de vezes em que $\mathrm{DM} = 0$. Esse é o procedimento sugerido pelos autores. No presente artigo, propõe-se um cálculo adicional referente ao desvio-padrão dos DM ao longo de $L$, o que servirá de fator de ponderação para o cálculo final do localizador MPG.

A Tabela 2 apresenta, em notação matricial, os DMs para um perfil extremo genérico, $k$. A penúltima linha da tabela representa a estatística de contagem, $\sum \mathrm{DM}_{kjl,r} = 0$, numa mesma execução $r$, ao longo das $L$ categorias, e a última linha corresponde ao desvio-padrão da distribuição de DMs ao longo de $L$. O cálculo do desvio-padrão é dado por:

$$(1)\quad \sigma(DM^{r}_{ljk}) = \sqrt{\frac{\left(DM^{r}_{lj} - \left(\sum_{j=1}^{p}\sum_{l=1}^{n} DM^{r}\right)^{-\sum_{j=1}^{p} l}\right)^{2}}{\sum_{j=1}^{p} l}}$$

$$\text{com } r = \{1,2,\cdots,m\}$$

Observe que o desvio-padrão é obtido para cada execução $r$, ao longo das $L$ categorias das $J$ variáveis, para um dado perfil $k$. O indicador $\sum \mathrm{DM}_{kjl,r} = 0$ deve ser ponderado pelo desvio-padrão sugerido.

A estatística proposta neste trabalho apresenta-se da seguinte forma:

$$*\frac{\sigma(DM^{r}_{kjl})}{\left(\sum_{j=1}^{p}\sum_{l=1}^{n} \#\, DM^{r}_{kjl} = 0\right)} \text{ com } r = \{1,2,\cdots,m\}$$

O numerador representa o desvio-padrão dos desvios médios, enquanto o denominador corresponde à contagem do número de desvios médios iguais a zero. Nesse sentido, a ponderação segue a lógica de uma estatística conhecida: o coeficiente de variação, que é representado pela divisão do desvio-padrão pela média. Quanto menor o valor desta estatística, mais estável e homogêneo é o conjunto de dados.

Uma vez ordenados os $\sum \#\,\mathrm{DM}_{kjl,r} = 0$ ponderados, para cada perfil $k$, em ordem crescente, chega-se a um MGP, no qual o peso é representado pela variabilidade dos desvios em relação à média das probabilidades. Cada perfil extremo, como alertado pelos autores, tem um MGP específico. Assim, o MGP final é obtido por meio da média dos MGP de cada um dos perfis:

$MGP_F = \left( \sum_{k=1}^{K} r'_{MGP_k} \right)^{-K}$, em que $r'$ corresponde à rodada aleatória de maior posicionamento em cada um dos perfis extremos $k$.

Deve-se considerar, ainda, a situação em que não se observa nenhum DM igual a zero. Nesse caso, especificamente, a estatística de MGP não pode ser calculada, uma vez que esta depende da contagem de desvios médios iguais a zero no denominador. Nesse contexto específico, sugere-se a utilização do próprio coeficiente de variação (CV), que é calculado por meio da razão entre o desvio-padrão e a média dos desvios médios:

$$CV = \frac{\sigma(DM^{r}_{kjl})}{\left( \sum_{l=1}^{n} DM^{r}_{kjl} \right)^{-n}} \quad \text{com} \quad r = \{1, 2, \cdots, m\}$$

Nesse contexto de ausência de DM iguais a zero, essa é uma estatística mais confiável do que o desvio-padrão apenas, uma vez que leva em consideração dois parâmetros (média e desvio-padrão).

**TABELA 1**
**Probabilidades de pertencimento ao perfil extremo *k* (lambdas), média dos lambdas por categoria *l* da variável *j* e desvio-médio por categoria *l* da variável *j* do perfil K=k num modelo gravitacional de K perfis extremos**

| Variável (j) | Categoria (l) | Execução (r) | | | | Média $\left( \sum_{r=1}^{m} \lambda^{r}_{l_n j_p k} \right)^{-m}$ |
|---|---|---|---|---|---|---|
| | | $r_1$ | $r_2$ | | $r_m$ | |
| 1 | 1 | $\lambda^{r_1}_{l_1 j_1 k}$ | $\lambda^{r_2}_{l_1 jk}$ | ... | $\lambda^{r_m}_{l_1 j_1 k}$ | $\left( \sum_{r=1}^{m} \lambda^{r}_{l_1 j_1 k} \right)^{-m}$ |
| | 2 | $\lambda^{r_1}_{l_2 j_1 k}$ | $\lambda^{r_2}_{l_2 jk}$ | ... | $\lambda^{r_m}_{l_1 j_1 k}$ | $\left( \sum_{r=1}^{m} \lambda^{r}_{l_2 j_1 k} \right)^{-m}$ |
| | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| | n | $\lambda^{r_1}_{l_n j_1 k}$ | $\lambda^{r_2}_{l_n jk}$ | ... | $\lambda^{r_m}_{l_1 j_1 k}$ | $\left( \sum_{r=1}^{m} \lambda^{r}_{l_n j_1 k} \right)^{-m}$ |
| 2 | 1 | $\lambda^{r_1}_{l_1 j_2 k}$ | $\lambda^{r_2}_{l_1 j_1 k}$ | ... | $\lambda^{r_m}_{l_1 j_1 k}$ | $\left( \sum_{r=1}^{m} \lambda^{r}_{l_1 j_2 k} \right)^{-m}$ |
| | 2 | $\lambda^{r_1}_{l_2 j_2 k}$ | $\lambda^{r_2}_{l_2 j_2 k}$ | ... | $\lambda^{r_m}_{l_2 j_2 k}$ | $\left( \sum_{r=1}^{m} \lambda^{r}_{l_2 j_2 k} \right)^{-m}$ |
| | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| | n | $\lambda^{r_1}_{l_n j_2 k}$ | $\lambda^{r_2}_{l_n j_2 k}$ | ... | $\lambda^{r_m}_{l_n j_2 k}$ | $\left( \sum_{r=1}^{m} \lambda^{r}_{l_n j_2 k} \right)^{-m}$ |
| p | 1 | $\lambda^{r_1}_{l_1 j_p k}$ | $\lambda^{r_2}_{l_1 j_p k}$ | ... | $\lambda^{r_m}_{l_1 j_p k}$ | $\left( \sum_{r=1}^{m} \lambda^{r}_{l_1 j_p k} \right)^{-m}$ |
| | 2 | $\lambda^{r_1}_{l_2 j_p k}$ | $\lambda^{r_2}_{l_2 j_p k}$ | ... | $\lambda^{r_m}_{l_2 j_p k}$ | $\left( \sum_{r=1}^{m} \lambda^{r}_{l_2 j_p k} \right)^{-m}$ |
| | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| | n | $\lambda^{r_1}_{l_n j_p k}$ | $\lambda^{r_2}_{l_n j_p k}$ | ... | $\lambda^{r_m}_{l_n j_p k}$ | $\left( \sum_{r=1}^{m} \lambda^{r}_{l_n j_p k} \right)^{-m}$ |

Fonte: Elaboração dos autores.

**TABELA 2**
**Desvios em relação à média (DM) das probabilidades de pertencimento ao perfil extremo *k* (lambdas) por categoria *l* da variável *j* do perfil K=k, máximo de DM nulos por execução r e desvio-padrão univariado dos DM por execução *r* num modelo gravitacional de *K* perfis extremos**

| Variável (j) | Categoria (l) | Execução (r) | | | |
|---|---|---|---|---|---|
| | | $r_1$ | $r_2$ | | $r_m$ |
| 1 | 1 | $DM^{r_1}_{l_1 j_1 k}$ | $DM^{r_2}_{l_1 jk}$ | ... | $DM^{r_m}_{l_1 j_1 k}$ |
| | 2 | $DM^{r_1}_{l_2 j_1 k}$ | $DM^{r_2}_{l_2 jk}$ | ... | $DM^{r_m}_{l_1 j_1 k}$ |
| | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| | n | $DM^{r_1}_{l_n j_1 k}$ | $DM^{r_2}_{l_n jk}$ | ... | $DM^{r_m}_{l_1 j_1 k}$ |
| 2 | 1 | $DM^{r_1}_{l_1 j_2 k}$ | $DM^{r_2}_{l_1 j_1 k}$ | ... | $DM^{r_m}_{l_1 j_1 k}$ |
| | 2 | $DM^{r_1}_{l_2 j_2 k}$ | $DM^{r_2}_{l_2 j_2 k}$ | ... | $DM^{r_m}_{l_2 j_2 k}$ |
| | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| | n | $DM^{r_1}_{l_n j_2 k}$ | $DM^{r_2}_{l_n j_2 k}$ | ... | $DM^{r_m}_{l_n j_2 k}$ |
| p | 1 | $DM^{r_1}_{l_1 j_p k}$ | $DM^{r_2}_{l_1 j_p k}$ | ... | $DM^{r_m}_{l_1 j_p k}$ |
| | 2 | $DM^{r_1}_{l_2 j_p k}$ | $DM^{r_2}_{l_2 j_p k}$ | ... | $DM^{r_m}_{l_2 j_p k}$ |
| | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| | n | $DM^{r_1}_{l_n j_p k}$ | $DM^{r_2}_{l_n j_p k}$ | ... | $DM^{r_m}_{l_n j_p k}$ |
| $\sum_{lj=11}^{np} DM^{r}_{ljk} = 0$ | | $\sum_{lj=11}^{np} DM^{r_1}_{ljk} = 0$ | $\sum_{lj=11}^{np} DM^{r_2}_{ljk} = 0$ | | $\sum_{lj=11}^{np} DM^{r_m}_{ljk} = 0$ |
| $\sigma\left(DM^{r}_{ljk}\right)$ | | $\sigma\left(DM^{r_1}_{ljk}\right)$ | $\sigma\left(DM^{r_2}_{ljk}\right)$ | | $\sigma\left(DM^{r_m}_{ljk}\right)$ |

Fonte: Elaboração dos autores.

## Comparando os localizadores do modelo de máximo global

*Base de dados*

Neste artigo, utilizou-se a mesma base de dados empregada por Guedes et al. (2010). A comparação baseada na mesma amostra analítica é importante devido à influência do nível de entropia presente na estrutura implícita aos dados sobre a capacidade do método GoM em localizar um máximo global (MANTON et al., 1994).

Assim, partiu-se das informações sobre uso e cobertura do solo, tamanho da propriedade, produção agrícola e estoque bovino (28 variáveis categóricas) para 293 lotes rurais residentes ao longo da Rodovia Transamazônica, em torno dos municípios de Altamira, Brasil Novo, Medicilândia e Uruará, no Estado do Pará. Os dados utilizados, coletados em 2005 e representativos dos lotes rurais da região de assentamento de Altamira, são parte do projeto *Amazonian Deforestation and the Structure of Households*, financiado pelo *National Institute of Child Health and Human Development* (NIH - HD35811-04), coordenado pelo investigador principal Dr. Emílio Moran e organizado por uma equipe de pesquisadores do *Anthropological Center for Training on Global Environmental Change* (ACT), na Indiana University. O projeto é uma parceria binacional e conta com a colaboração de pesquisadores do Núcleo de Estudos Populacionais (Nepo), da Universidade Estadual de Campinas.[1]

[1] Para mais detalhes, ver Moran et al. (2007).

Os dados empregados serviram de base para a caracterização de sistemas de uso do solo, considerando-se a escala de produção, o tipo de cultura e a destinação final da produção nos lotes rurais da região em torno de Altamira.[2]

*Resultados*

Neste artigo, a discussão restringe-se à capacidade do localizador MGP de reduzir a distância do modelo final à estrutura latente real, observada por meio de um menor valor de AIC (*Akaike Information Criterion*).

Para efeito de comparação, os resultados mostrados nesta seção são baseados no mesmo número de perfis extremos (três) utilizados no artigo de Guedes et al. (2010), evitando, assim, que se obtenham modelos finais com números distintos de parâmetros estimados (dimensionalidade). Apesar de os modelos apresentarem dimensionalidade fixa, o AIC pode variar em decorrência dos valores finais da função de verossimilhança, conforme sugere a fórmula da estatística de ajuste (AKAIKE, 1973):

$$AIC = 2p - 2\ln(L),$$

onde:

$p$ = número de parâmetros finais estimados ($\lambda_{kjl}$ e $g_{ik}$);

$L$ = valor convergente (máximo) da função de verossimilhança.

Um modelo com menor distância aos dados foi interpretado como aquele que apresenta o menor AIC. Neste artigo, tomou-se a redução do AIC como um indicador de melhoria no ajuste do modelo identificado baseado no localizador $MGP_F$.

A Tabela 3 apresenta o posicionamento dos 30 modelos com parâmetros estáveis (seguindo procedimento de estabilização descrito por Guedes et al., 2010), para ambos os localizadores empíricos: MG e $MGP_F$. Os resultados indicam que o modelo de máximo global identificado pelo localizador MG corresponde à execução aleatória R05. Quando utilizado o $MGP_F$, o modelo ótimo desloca-se para a execução aleatória R13. Observando o valor de AIC para ambos os modelos finais, percebe-se que $AIC_{MG}$ (15.259,54) > $AIC_{MGP}$ (15.096,76).

Tomando as diferenças nos valores de AIC para todas as execuções aleatórias de mesma posição e executando um teste de médias, obteve-se uma diferença estatisticamente significativa a menos de 5% (valor de $p = 0{,}0484$). Nossa hipótese nula é a de que a média das diferenças do AIC = 0 para as cinco primeiras posições, ou seja, inexistência de ganho quantificável na proximidade do modelo à estrutura implícita com base em quaisquer dos localizadores empregados:

Hipótese nula ($H_0$):

$$AIC^{r}_{MGP} - AIC^{r}_{MG} = 0 \quad \text{com} \quad r = \{1, 2, \cdots, m\}$$

Hipótese alternativa ($H_A$):

$$AIC^{r}_{MGP} - AIC^{r}_{MG} < 0 \quad \text{com} \quad r = \{1, 2, \cdots, m\}$$

O resultado do teste foi interpretado como um indicador de redução das distâncias médias dos parâmetros estruturais obtidos pelo modelo localizado por $MGP_F$, se comparado às distâncias geradas pelo modelo identificado por MG. Os resultados em conjunto sugerem que desconsiderar a variabilidade na localização do modelo de máximo global pode afetar significativamente a distância dos dados à centralidade da amostra e causar viés de consistência nos estimadores finais.

## Considerações finais

O modelo GoM tem sido amplamente utilizado, especialmente na área de Demografia. Este trabalho avança ao indicar um procedimento adicional em busca de um modelo com parâmetros, que melhor descreva os dados.

Com o objetivo de encontrar esse modelo, estudos recentes indicam que o ideal é efetuar várias execuções com matriz de probabilidades iniciais aleatórias. Ao serem obtidas várias matrizes de probabilidades finais, seria possível ao pesquisador observar uma convergência em torno de certos valores de probabilidades recorrentes. Guedes et al. (2010) sugeriram um

[2] Mais detalhes sobre os sistemas gerados encontram-se em Guedes (2010).

**TABELA 3**
**Posição dos modelos de ordem K=3, AIC por localizador empírico do máximo global e diferença entre AICs – Sistemas de Uso do Solo na Região de Estudo de Altamira, Pará (N = 293 lotes rurais)**

| Posição (ranking) | Rodada aleatória (r) | | AIC | | $AIC_{MGP}$ - $AIC_{MG}$ |
|---|---|---|---|---|---|
| | MG | MGP | MG | MGP | |
| 1 | R05 | R13 | 15259,54 | 15096,76 | -162,79 |
| 2 | R01 | R25 | 15110,01 | 15100,27 | -9,74 |
| 3 | R20 | R22 | 15442,39 | 15354,17 | -88,22 |
| 4 | R09 | R05 | 15261,67 | 15259,54 | -2,13 |
| 5 | R22 | R12 | 15354,17 | 15097,97 | -256,20 |
| 6 | R04 | R29 | 15096,54 | 15110,78 | 14,23 |
| 7 | R29 | R20 | 15110,78 | 15442,39 | 331,61 |
| 8 | R30 | R02 | 15111,03 | 15099,59 | -11,44 |
| 9 | R14 | R06 | 15095,34 | 15099,61 | 4,27 |
| 10 | R07 | R30 | 15594,07 | 15111,03 | -483,04 |
| 11 | R08 | R09 | 15101,09 | 15261,67 | 160,58 |
| 12 | R26 | R23 | 15107,71 | 15095,19 | -12,51 |
| 13 | R21 | R24 | 15427,42 | 15101,68 | -325,74 |
| 14 | R28 | R08 | 15285,71 | 15101,09 | -184,62 |
| 15 | R11 | R14 | 15097,16 | 15095,34 | -1,82 |
| 16 | R18 | R07 | 15308,99 | 15594,07 | 285,08 |
| 17 | R23 | R01 | 15095,19 | 15110,01 | 14,82 |
| 18 | R02 | R04 | 15099,59 | 15096,54 | -3,05 |
| 19 | R03 | R18 | 15110,54 | 15308,99 | 198,45 |
| 20 | R12 | R21 | 15097,97 | 15427,42 | 329,45 |
| 21 | R06 | R26 | 15099,61 | 15107,71 | 8,09 |
| 22 | R10 | R28 | 15110,22 | 15285,71 | 175,50 |
| 23 | R13 | R17 | 15096,76 | 15106,68 | 9,93 |
| 24 | R19 | R27 | 15098,59 | 15099,24 | 0,65 |
| 25 | R25 | R10 | 15100,27 | 15110,22 | 9,95 |
| 26 | R27 | R19 | 15099,24 | 15098,59 | -0,65 |
| 27 | R24 | R03 | 15101,68 | 15110,54 | 8,86 |
| 28 | R17 | R11 | 15106,68 | 15097,16 | -9,52 |
| 29 | R15 | R15 | 15101,05 | 15101,05 | 0,00 |
| 30 | R16 | R16 | 15114,36 | 15114,36 | 0,00 |
| Média (5) | - | - | - | - | -103,814 |

Fonte: Anthropological Center for Training on Global Environmental Change (2010); Guedes et. al. (2010).

procedimento de localização da execução mais informativa, com base em uma estatística de desvio médio, obtida de várias execuções e da média das probabilidades destas execuções. Desvios em torno da probabilidade média iguais a zero seriam indicativos de que não haveria grandes diferenças de uma dada execução, para cada perfil, em relação a cada probabilidade média (máximo global, ou MG). Contudo, esta estatística não incorporava qualquer medida de variabilidade. Neste artigo, essa limitação foi contornada a partir da utilização de uma estatística de ponderação semelhante ao coeficiente de variação (Média Global Ponderada, ou MGP). Esse indicador buscou não penalizar de forma desproporcional situações nas quais os desvios médios, apesar de diferentes de zero, são muito pequenos. Os resultados de ajuste do modelo, dados pelo critério de Akaike, revelaram que os achados obtidos

por intermédio do MGP foram melhores do que aqueles alcançados pelo MG.

A redução do AIC sugere que localizadores de modelos que consideram a variabilidade nos desvios em relação à média das probabilidades estimadas aumentam, com maior confiabilidade, a proximidade da verdadeira estrutura latente aos dados.

**Referências**


AKAIKE, H. Information theory and an extension of the maximum likelihood principle. In: PETROV, B. N.; CSAKI, F. (Eds.). **Second International Symposium on Information Theory**. Budapest: Akademia Kiado, 1973, p. 267-281.

ALVES, L. C.; LEITE, I. C.; MACHADO, C. J. Perfis de saúde dos idosos no Brasil: análise da Pesquisa Nacional por Amostra de Domicílios de 2003 utilizando o método *Grade of Membership*. **Cadernos de Saúde Pública**, v. 24, n. 3, p. 535-546, 2008.

CAETANO, A. J.; MACHADO, C. J. Consistência e identificabilidade no modelo *Grade of Membership*: uma nota metodológica. **Revista Brasileira de Estudos de População**, v. 26, n. 1, p. 145-149, 2009.

DRUMOND, E. F.; MACHADO, C. J.; FRANCA, E. Óbitos neonatais precoces: análise de causas múltiplas de morte pelo método *Grade of Membership*. **Cadernos de Saúde Pública**, v. 23, n. 1, p. 157-166, 2007.

GARCIA, R. A.; SOARES-FILHO, B. S.; SAWYER, D. O. Socioeconomic dimensions, migration, and deforestation: an integrated model of territorial organization for the Brazilian Amazon. **Ecological Indicators**, v. 7, n. 3, p. 719-730, 2007.

GILES, R. The concept of grade of membership. **Fuzzy Sets and Systems**, v. 25, n. 3, p. 297-323, 1988.

GUEDES, G. R. **Ciclo de vida domiciliar, ciclo do lote e dinâmica do uso da terra na Amazônia rural brasileira** – Um estudo de caso para Altamira, Pará. Tese (Doutorado). Belo Horizonte, Cedeplar/UFMG, 2010.

GUEDES, G. R.; CAETANO, A. J.; MACHADO, C. J.; BRONDIZIO, E. S. Identificabilidade e estabilidade dos parâmetros no método *Grade of Membership* (GoM): considerações metodológicas e práticas. **Revista Brasileira de Estudos de População**, v. 27, n. 1, 2010.

GUEDES, G. R.; COSTA, S. M.; BRONDIZIO, E. S. Hierarchy of urban areas in the brazilian Amazon and its environmental implications. **UGEC Viewpoints**, n.2, p. 25-27, 2009a.

_____. Revisiting the hierarchy of urban areas in the brazilian Amazon: a multilevel approach. **Population & Environment**, v. 30, p. 159–192, 2009b.

GUEDES, G. R.; RESENDE, A. C.; BRONDIZIO, E. S.; PENNA-FIRME, R. P.; CAVALLINI, I. Poverty dynamics and income inequality in the eastern brazilian Amazon: a multidimensional approach. In: XXVI IUSSP CONFERENCE. **Anais...** Marrakesh, Marrocos, 2009c.

KAUFMAN, L.; ROUSSEEUW, P. J. **Finding groups in data: an introduction to cluster analysis**. New York: John Wiley, 1990.

MANTON, K. G.; WOODBURY, M. A.; TOLLEY, H. D. **Statistical application using fuzzy sets**. Nova York: John Wiley & Sons, 1994.

MELO, F. L. B.. Casais na Grande São Paulo: investigando a diversidade. **Nova Economia**, v. 17, n. 2, p. 207-240, 2007.

MIRANDA-RIBEIRO, P.; SIMÃO, A. B.; CAETANO, A. J.; PERPÉTUO, I. H. O.; LACERDA, M. A.; TORRES, M. E. A. Acesso à contracepção e ao diagnóstico do câncer de colo uterino em Belo Horizonte: uma contribuição metodológica aos estudos quanti-quali. **Revista Brasileira de Estudos de População**, v. 24, p. 341-344, 2007.

MORAN, E. F.; VANWEY, L. K.; CARMO, R.; HOGAN, D. **Amazonian deforestation and the structure of households (phase III)**. Grant Proposal sponsored by the National Institutes of Child Health and Human Development, jul. 2007. 37p. (Grant # 2R56HD035811-08,

NIH, IRG: ZRG1). Disponível em: <http://www.researchgrantdatabase.com/g/2R01HD035811-04/Amazonian-Deforestation-and-the-Structure-of-Households/>. Acesso em: 08 out. 2008.

PEARCE, L. D. Integrating survey and ethnographic methods for systematic anomalous case analysis. **Sociological Methodology**, v. 32, p. 103-132, 2002.

SANTOS, M. **A influência da dinâmica demográfica e domiciliar no processo de ocupação do Cerrado brasileiro**: o caso do Programa de Assentamento Dirigido do Alto Paranaíba, Minas Gerais, Brasil. Tese (Doutorado). Belo Horizonte: Centro de Desenvolvimento e Planejamento Regional – Cedeplar/UFMG, 2010.

SAWYER, D. O; LEITE, I. C.; ALEXANDRINO, R. Perfis de utilização de serviços de saúde no Brasil. **Ciência e Saúde Coletiva**, v. 7, n. 4, p. 757-776, 2002.

VELOSO, A. A.; SIQUEIRA, G. M.; PÔSSAS, B. A. V. E.; MEIRA JUNIOR, W.; CARVALHO, M. L. B. Mineração incremental de regras de associação. In: XVI SBBD – SIMPÓSIO BRASILEIRO DE BANCO DE DADOS. **Anais...** Rio de Janeiro, 2001.

WOODBURY, M. A.; CLIVE, J. Clinical pure types as a fuzzy partition. **Journal of Cybernetics and Systems**, v. 4, n. 3, p. 111-121, 1974.

**Resumen**

*Incorporando la variabilidad en el proceso de identificación del modelo de máximo global en el Grade of Membership (GoM): consideraciones metodológicas*

La disponibilidad de bases de datos cada vez más complejas y multidimensionales es uno de los principales factores motivadores para el aumento del número de estudios que utilizan análisis multivariados basados en la lógica de conjuntos nebulosos. A pesar de la diseminación del método Grade of Membership en los trabajos empíricos brasileños dentro del área de ciencias sociales y salud, cuestiones relativas a la identificabilidad y estabilidad de los parámetros finales, estimados por el programa GoM 3.4, no fueron suficientemente profundizadas. Dada la relevancia de que se obtengan parámetros únicos y estables, Guedes et al. (2010) propusieron un procedimiento empírico para localizar un modelo de máximo global (MG) con parámetros estables. No obstante, su localizador de MG no incorpora cualquier medida de variabilidad. En este artículo, tal limitación se sortea mediante la utilización de una estadística de ponderación –Máximo Global Ponderado (MGP)- semejante al coeficiente de variación. Este indicador busca no penalizar de forma desproporcionada situaciones en las que los desvíos medios, a pesar de ser diferentes a cero, son muy pequeños. Se presentan evidencias de que el localizador MGP reduce la distancia del modelo identificado respecto a la estructura real latente de los datos en análisis, cuando se comparan con el modelo identificado por el localizador no ponderado, MG.

**Palabras-clave:** *Grade of Membership*. Máximo Global Ponderado. Variabilidad. Identificabilidad.

**Abstract**

*Incorporating variability in the process of identification of the global maximum model in Grade of Membership (GoM): methodological considerations*

The availability of increasingly complex and multidimensional datasets is one of the main causes for the increase in studies employing multivariate analyses based on fuzzy sets. Even though the Grade of Membership method has been widely used in Brazil for empirical studies in health and social sciences, issues regarding identifiability and stability of the final parameters

estimated by GoM 3.4 software have not been thoroughly examined. Given the relevance of unique and stable parameters, Guedes et al. (2010) proposed an empirical method to locate a global maximum (GM) with stable parameters. However, the GM locator does not incorporate variability. In the present article, this limitation is circumvented by employing a weighted statistic – weight global maximum (WGM) – similar to the variation coefficient. This indicator does not affect disproportionately situations with very low mean deviations. The WGM locator is shown to decrease the distance of the identified model from the real structure, when compared with the GM locator.

**Keywords:** *Grade of Membership*. Weighted Global Maximum. Variability. Identifiability.

Recebido para publicação em 03/05/2010
Aceito para publicação em 06/08/2010