# Identificabilidade e estabilidade dos parâmetros no método *Grade of Membership* (GoM): considerações metodológicas e práticas

**Gilvan Ramalho Guedes**[*]
**André Junqueira Caetano**[**]
**Carla Jorge Machado**[***]
**Eduardo Sonewend Brondízio**[****]

*O método Grade of Membership (GoM) tem sido cada vez mais utilizado por demógrafos brasileiros e tem a vantagem de possuir um parâmetro que mensura a heterogeneidade individual, com base nas correlações não-observáveis entre as categorias de resposta das variáveis de interesse, gerando um medida do grau de pertencimento de cada indivíduo a perfis extremos. Alguns autores, contudo, chamam atenção para questões importantes na calibragem dos modelos finais que utilizam o programa GoM versão 3.4, como o problema de identificabilidade – soluções múltiplas para parâmetros estimados. Neste artigo, é sugerido um procedimento capaz de identificar um modelo final com solução única que descreva os tipos puros mais fidedignos à base de dados, em uma tentativa de otimização. Para ilustrar esse processo, utilizou-se uma base de dados correspondente a um levantamento econômico e sociodemográfico de uma população de pequenos agricultores residentes ao longo da Rodovia Transamazônica, no Estado do Pará. Também identificou-se a existência de instabilidade nos parâmetros estimados pelo programa GoM 3.4, sendo proposto um método de estabilização de seus valores. Com esses procedimentos combinados, os usuários do programa GoM 3.4 poderão descrever sua base de dados de forma mais adequada e responder às críticas sobre questões de identificabilidade e estabilidade dos modelos resultantes. Essas soluções empíricas são relevantes por afetarem cálculos de prevalência e de incidência de eventos de interesse, além de trazerem consequências importantes sobre o ponto e o momento corretos para intervenções de políticas públicas ou de planejamento prospectivo em análises de projeção.*

**Palavras-chave**: *Grade of Membership*. Identificabilidade. Estabilidade. Máximo global. Conjuntos nebulosos.

## Introdução

A necessidade de procurar padrões frequentes e extrair grupamentos em bancos de dados se faz presente em muitas áreas do conhecimento. O rápido crescimento da complexidade, multidimensionalidade e quantidade de dados em todas as áreas, bem como a necessidade de extrair informações úteis de dados coletados, é a motivação básica para a procura de algoritmos variados para a mineração de dados (*data mining*)

[*] Doutor em Demografia, pesquisador residente do Environmental Change Initiative / Brown University.
[**] Ph.D in Sociology, professor adjunto III da Pontifícia Universidade Católica de Minas Gerais.
[***] Ph.D in Population Dynamics, professora adjunta III da Universidade Federal de Minas Gerais.
[****] Ph.D in Anthropology, professor de antropologia, professor adjunto de Ciências Ambientais e chefe de Departamento na Indiana University.

com a finalidade de descobrir conhecimentos (*knowledge discovery database*) implícitos em bases de dados (VELOSO et al., 2001).

Do mesmo modo, o uso de tipologias para caracterização e categorização social tem sido historicamente uma ferramenta analítica útil, porém controversa, nas ciências sociais. Tipologias, quantitativas ou qualitativas, permitem simplificar e generalizar um determinado *continuum*, embora promovam categorias estanques que podem mascarar a heterogeneidade interna de grupos. O método *Grade of Membership* (GoM), ao parametrizar a heterogeneidade individual, tende a superar a necessidade de criação de tipologias arbitrárias, reduzindo, portanto, os conteúdos implícitos. Ao contrário, os parâmetros representativos dos graus de pertencimento individual aos perfis extremos expandem as associações implícitas ao dado, uma vez que consideram essas associações no nível da categoria de cada variável, e não entre os indivíduos, como nas técnicas de agrupamento baseadas em lógica binária (MANTON et al., 1994). Ou seja, assume-se que a variação ocorre entre os indivíduos e que não é meramente aleatória.

O método GoM vem sendo amplamente utilizado na demografia latino-americana, em especial entre os demógrafos brasileiros (SAWYER et al., 2002; DRUMOND et al., 2007; MELO, 2007; ALVES et al., 2008; GUEDES et al., 2009a, 2009b, 2009c). Os trabalhos que utilizam o GoM têm em comum contextos em que encontrar estruturas implícitas nos dados é essencial, ou seja, estruturas que revelem os padrões de ocorrências conjuntas de valores específicos de variáveis – grupamentos (MIRANDA-RIBEIRO et al., 2007; GUIMARÃES et al., 2008). O modelo, além de permitir que sejam definidos padrões – chamados *perfis extremos* – capazes de sintetizar grande parte da informação contida na base de dados para os indivíduos que a compõem, também possibilita a avaliação das proximidades – medidas por escores de grau de pertencimento – dos indivíduos a cada um dos perfis extremos (GILES, 1988; MANTON et al., 1994). Um perfil extremo é definido apenas para os indivíduos cujos escores a esse perfil sejam iguais a 1 (indivíduos com total pertencimento, chamados tipos puros), sendo caracterizado por um conjunto de probabilidades de resposta *l* do indivíduo *i* (tipo puro) ao perfil *k* na variável *j*, $\lambda_{kjl}$.

Cabe observar (e justificar) que, embora os perfis possam sintetizar grande parte da informação contida para os indivíduos, muitos deles podem, naturalmente, apresentar características de mais de um perfil, em decorrência da heterogeneidade existente nas populações. O GoM utiliza um procedimento iterativo, que busca a convergência de resultados, mas depende de uma matriz de probabilidades iniciais ($\lambda_{kjl}$ iniciais) como insumo para que o algoritmo possa ser executado. Consequentemente, dependendo da matriz inicial de valores fornecida pelo pesquisador[1] ou gerada pelo programa (aleatoriamente ou de alguma outra forma especificada[2]), os resultados finais para os parâmetros estimados podem variar em sucessivas execuções.

Essa constatação faz emergir uma preocupação natural do pesquisador interessado em encontrar uma descrição "correta" e fidedigna de seus dados: obter um modelo

[1] O fornecimento de uma matriz de probabilidades iniciais pode ser derivado de instrumentos qualitativos, para minimizar a chance de se obter um modelo de máximo local. A matriz de probabilidade pode ser informada por técnicas de entrevistas semiestruturadas ou resultante de discussões levantadas por grupos focais, baseando-se nas variáveis de interesse. Nesse caso, espera-se que as probabilidades iniciais sejam direcionadas por prevalências obtidas empiricamente.

[2] Há outras formas de definição da matriz inicial de probabilidades. Um procedimento útil, em especial quando se deseja estabelecer perfis que guardem entre si uma estrutura de hierarquia, é definir que a matriz seja gerada por PURE1, disponível no programa GoM versão 3.4. Com esse procedimento, os componentes do perfil extremo 1, num modelo de K perfis, terá os valores mais baixos das categorias das J variáveis internas utilizadas na definição do modelo final (GUEDES et al., 2009c). Finalmente, ressalte-se que há outros programas, como o DSI GOM (Decision System Inc. s.d), que utilizam outros procedimentos para dar início ao processo de convergência. No caso do DSI GOM, são muitas as restrições impostas pelo programa, pois tanto a matriz de probabilidades iniciais quanto o número de perfis são condicionados por uma variável denominada variável indicadora, que é previamente definida pelo pesquisador.

que seja *identificável*, isto é, com uma única solução. Com efeito, ao se empregar o procedimento de seleção aleatória para os primeiros $\lambda_{kjl}$ – probabilidades representativas dos perfis extremos –, pode-se chegar a resultados que correspondem a máximos locais, em vez de máximos globais (CAETANO; MACHADO, 2009). Isso ocorre porque o processo iterativo utilizado pelo algoritmo do programa GoM versão 3.4 não garante, por si só, a obtenção de perfis extremos que representem de forma ótima os tipos puros reais presentes na amostra.

Em algumas circunstâncias, no entanto, dado um modelo de *K* perfis, a mudança na localização de um perfil extremo de sua posição k = 1 para k = 3, por exemplo, ocorre independentemente da questão da identificabilidade. Como a matriz inicial de probabilidades pode ser definida de modo aleatório, é possível um perfil extremo *k* em uma execução *r* qualquer estar localizado em outro $k = k'$, quando analisado um modelo distinto com mesmo *K*, estimado em uma execução $r = r'$. Esse reposicionamento ocorre com muita frequência ao longo das execuções.

Segundo Guedes et al. (2009a), durante a classificação da hierarquia urbana na Amazônia, Belém, frequentemente modificava sua posição nos perfis extremos. Trabalhando com um modelo de três perfis extremos, na maioria das execuções o terceiro perfil era o que concatenava as áreas urbanas municipais de maior hierarquia. Em um número não desprezível de execuções, no entanto, a capital do Pará e todas as demais áreas urbanas municipais correlatas passavam a pertencer ao perfil extremo 2 ou 1. Assim, o perfil extremo 3 deixava, para aquela execução, de incluir as áreas urbanas de maior posição hierárquica.

O problema da identificabilidade, portanto, não tem relação com a localização do perfil extremo em sucessivas execuções, mas refere-se à dificuldade de se encontrar um perfil extremo que, independente da sua localização (do seu *k* em um modelo de *K* perfis), represente uma solução única que descreva as características definidoras dos tipos puros "reais".

Em adição ao reposicionamento dos perfis extremos em sucessivas execuções aleatórias, o problema da convergência parcial, como será visto neste trabalho, interfere não somente na identificabilidade, mas também na *estabilidade* dos parâmetros estimados pelo GoM. Assim, qualquer aplicação empírica do modelo GoM deve ser capaz de atender a essas duas propriedades: identificabilidade e estabilidade estrutural. Neste estudo, procurou-se avançar a questão de identificabilidade do modelo – iniciada por Caetano e Machado (2009) –, utilizando um procedimento operacionalmente simples que sugere a localização empírica do modelo de máximo global. Também é sugerida uma rotina que estabiliza os parâmetros estimados, solucionando a questão da instabilidade desses. Tais procedimentos combinados procuram facilitar, ao usuário final, a seleção da melhor execução que descreva seus dados.

## Em busca de uma medida de identificabilidade do modelo de máximo global

O algoritmo[3] utilizado no programa GoM, versão 3.4, baseado em processo iterativo, gera dois problemas empíricos principais: a identificabilidade do modelo não é garantida (CAETANO; MACHADO, 2009); e existe instabilidade estrutural dos parâmetros finais estimados. A identificabilidade refere-se à estimação do modelo que melhor descreva tanto os perfis extremos (conjunto de $\lambda_{kjl}$) quanto a heterogeneidade presente nos dados ($g_{ik}$). Quanto à identificabilidade, os parâmetros ($g_{ik}$ e $\lambda_{kjl}$) deveriam ter solução única, uma vez que, segundo Manton et al. (1994), os perfis extremos definidos com base em um conjunto convexo com a menor dimensionalidade capaz de incorporar toda a densidade de probabilidade são vértices *únicos* e *fixos* no espaço convexo (*simplex*). Na prática, no entanto, os modelos finais em sucessivas execuções variam, descrevendo vértices

[3] O algoritmo utilizado na versão 3.4 do programa GoM foi proposto por Woodbury e Clive (1974).

não-estáveis, levando a aparentes máximos, ou máximos locais (não globais). O máximo global, portanto, deve representar, de alguma forma, os vértices mais estáveis e que melhor descrevam a heterogeneidade total da amostra. A instabilidade dos parâmetros, por seu turno, está associada à sua não-convergência aos valores estáveis após a primeira solução para o máximo da função de verossimilhança, mais detalhada a seguir.

## Dados e procedimento para verificação empírica da convergência e estabilidade

Neste trabalho, sugere-se um procedimento operacional para que um modelo de máximo global possa ser identificado entre diversos modelos gerados, tendo como ponto de partida uma mesma base de dados. Para tanto, utilizou-se uma base de dados com informações sobre classes de uso/cobertura do solo, estoque de gado e produção agrícola entre pequenos agricultores residentes no entorno das cidades de Altamira, Brasil Novo, Medicilândia e Uruará, no Estado do Pará (GUEDES et al., 2009d; VANWEY et al., 2008). Os dados referem-se a 2005 e a amostra selecionada com informações válidas totalizou 293 lotes rurais caracterizados por 28 variáveis.

Seguindo sugestão operacional de Caetano e Machado (2009), foram efetuadas aproximadamente 30 execuções com seleção aleatória dos primeiros $\lambda_{kjl}$ (a matriz inicial de probabilidades utilizadas como valores de entrada durante o processo iterativo). Como existe o problema de identificabilidade, efetuaram-se 30 execuções aleatórias para K = 2, K = 3, K = 4, K = 5 e, somente após a obtenção dos máximos globais para cada modelo de *K* variando de 1 a k perfis, calculou-se a estatística AIC (*Akaike Information Criterion*) (AKAIKE, 1973) e compararam-se seus valores finais[4] (Tabela 1).

**TABELA 1**
**Valores do Critério de Informação de Akaike (AIC), segundo número de perfis extremos dos sistemas de uso do solo**
**Região de estudo (1) – 2005**

| Número de perfis extremos | Número de parâmetros | Valor do ln(L)[II] | AIC[I] |
|---|---|---|---|
| 2 | 778 | -6857,13 | 15270,3 |
| 3 | 1167 | -6462,77 | 15259,5 |
| 4 | 1556 | -6101,54 | 15315,1 |
| 5 | 1945 | -5927,62 | 15745,2 |

Fonte: Dados de *survey* conduzido em Altamira (2005).
(1) Compreende o entorno das cidades de Altamira, Brasil Novo, Medicilândia e Uruará, no Estado do Pará. Nota: Fórmula do AIC = 2p – 2ln(L). L = função de máxima verossimilhança.

[4] Na verdade, o cálculo do AIC para seleção final do modelo com o melhor número de perfis extremos foi efetuado somente após a identificação do máximo global com estabilidade dos parâmetros (implementando o procedimento sugerido mais adiante, de autoalimentação dos valores de convergência dos $\lambda_{kjl}$ como valores iniciais a cada nova execução, até que a variação entre um $\lambda_{kjl}$ de uma execução anterior e da seguinte fosse nula entre todas as estimativas, $\lambda_{kjl}$, ao longo de todos os k perfis extremos). A seleção desse modelo final ao longo de vários K não é abordada aqui, trata-se do problema de seleção para K fixo.

O procedimento para identificação quantitativa do máximo global sugerida neste trabalho é o seguinte:

- efetuar de 20 a 30 execuções utilizando a matriz aleatória de parâmetros iniciais de e $\lambda_{kjl}$ e $g_{ik}$;
- essas execuções aleatórias devem ser realizadas para vários modelos com K variando de 2 a aproximadamente 5 perfis extremos, ou até que o AIC atinja o ponto mínimo. Por exemplo, se $AIC_{K=4}>AIC_{K=5}$, deve-se tentar identificar um modelo com K = 6 e observar se $AIC_{K=6}>AIC_{K=5}$. Na prática, o AIC mínimo é encontrado antes de K = 5 (CAETANO; MACHADO, 2009).[5] Neste trabalho, utilizou-se um modelo empírico para efeito ilustrativo, no qual o ponto de AIC mínimo ocorreu com K = 3, ou seja, um modelo com três perfis extremos. A regra geral é utilizar o modelo $AIC_K$ que atenda à restrição: $AIC_{K-1}>AIC_K<AIC_{K+1}$;
- para cada execução com matriz inicial aleatória fixando K perfis, obtém-se o número de parâmetros $\lambda_{kjl}$ igual a K multiplicado por J multiplicado por L (K*J*L). Por exemplo, em um modelo com L = 4, J = 30 e K = 3, tem-se um total de 360 estimativas de $\lambda_{kjl}$;
- em r execuções aleatórias, para uma mesma categoria l, de uma variável j pertencente a um mesmo perfil *k*, obtêm-se *r* probabilidades de resposta $\lambda_{kjl}$. Assim, é possível calcular a média dessas probabilidades obtidas ao longo das *r* execuções aleatórias, específica por categoria de uma variável em cada um dos perfis extremos e, então, obter uma *estatística de desvio em relação à média (DM)*, ao subtrair cada uma destas probabilidades a média da distribuição:

$$DM_{kjl,\,r}=\lambda_{kjl,\,r}-\frac{\sum_{r=1}^{30}\lambda_{kjl,\,r}}{r}$$

onde: $DM_{kjl,r}$ é o desvio da probabilidade estimada ($\lambda_{kjl}$) na *r*-ésima execução em relação à média das probabilidades em *r* execuções; $\lambda_{kjl}$ é a probabilidade de resposta *l* da variável *j* no perfil *k*, definida para os *k* tipos puros; *r* é o número de execuções;

- em situação de convergência para um mesmo valor em execuções aleatórias sucessivas, o somatório de $DM_{kjl,r}$, ao longo de todas as execuções, seria zero em módulo. Na prática, contudo, o número de vezes em que o desvio da média de uma categoria de uma variável em um perfil específico é igual a zero é sempre menor que *r*, resultando somatório não nulo. Para encontrar qual a posição dos desvios em termos de hierarquia do menor para o maior desvio médio a cada execução *r*, por perfil *k*, é possível estabelecer uma estatística de contagem ao longo das *l* categorias das variáveis *j*, ou seja, o número de vezes em que o desvio calculado para cada um dos *k* perfis é igual a zero. A estatística de desvio é contabilizada ao longo das categorias, *l*, por execução, e não mais ao longo das execuções;
- a condição anterior fornece uma distribuição de número de vezes em que o desvio médio é igual a zero para cada execução. É importante ressaltar que o *K* aqui corresponde ao número de perfis

[5] Isso nem sempre é verdade quando se utiliza o critério da razão de máxima verossimilhança. Vários estudos empíricos (CASSADY et al., 2001; WOODBURY; CLIVE, 1974), baseando-se no teste da razão de máxima verossimilhança, chegaram a modelos finais com 10 a 15 perfis extremos. A vantagem de utilizar a estatística AIC é que ela penaliza um modelo com mais perfis, pois considera o número de parâmetros, ao contrário do teste da razão de máxima verossimilhança, que se baseia somente no valor final de L entre dois modelos aninhados.

previamente definido. Ou seja, se K = 2, então tem-se o cálculo das estatísticas DM para k = 1 e k = 2. Importante lembrar que o valor de DM é influenciado por três fatores: número de execuções (*r*); relação ($\lambda_{kjl}$ - $\lambda_{kjl(médio)}$); e número de categorias (*l*);

- para cada conjunto de probabilidades $\lambda_{kjl}$, para um dado *k*, pode-se definir a classificação do número de vezes em que o desvio médio foi igual a zero (r = 1, ... , 30). Quanto maior o número de desvios nulos, maior a posição em termos de classificação para aquela execução *r* específica;
- no entanto, em um modelo com três perfis, por exemplo, k = 1, 2, 3, a execução aleatória de maior posição em termos de desvio médio pode diferir entre os perfis. Assim, para a obtenção do máximo global, é necessário que as 30 execuções (r = 1, ..., 30) sejam classificadas, em ordem crescente, para cada perfil (com 1 representando a melhor posição e 30 a inferior). A aplicação do procedimento aos sistemas de uso do solo, na região de Altamira, Pará (VANWEY et al., 2008) deu origem à classificação apresentada na Tabela 2, com k=1, 2, 3 perfis extremos definidos;
- a Tabela 2 explicita que, para cada perfil extremo, *k*, a execução aleatória, *r*, com o número máximo de desvios nulos difere. Por exemplo, para k = 1, r($DM_{máx}$) = R05; para k = 2, r($DM_{máx}$) = R09; e, por fim, para k = 3, r($DM_{máx}$) = R20. Para encontrar o melhor ajuste final ao longo dos *k* perfis, pode-se reordenar cada perfil *k* ascendentemente pela execução, *r*, e alcançar a média entre as posições obtidas nos *K* perfis (Tabela 3);
- para identificar o máximo global, portanto, basta selecionar a execução cuja classificação média foi menor (última coluna da Tabela 3). No exemplo, a execução contendo o máximo global, no caso de três perfis extremos, foi a R05.

**TABELA 2**
**Classificação da somatória de $DM\lambda_{kjl,r}$= 0 (# DM), por perfil e execução**
**Região de estudo (1) – 2005 (n=293)**

| PE1 | | | PE2 | | | PE3 | | |
|---|---|---|---|---|---|---|---|---|
| Rodada | #DM | Posição | Rodada | #DM | Posição | Rodada | #DM | Posição |
| R05 | 39 | 1 | R09 | 38 | 1 | R20 | 37 | 1 |
| R22 | 38 | 2 | R01 | 35 | 2 | R05 | 36 | 2 |
| R29 | 37 | 3 | R07 | 33 | 3 | R01 | 32 | 3 |
| R30 | 37 | 4 | R08 | 32 | 4 | R04 | 32 | 4 |
| R09 | 36 | 5 | R20 | 32 | 5 | R14 | 32 | 5 |
| R26 | 34 | 6 | R30 | 32 | 6 | R18 | 32 | 6 |
| R07 | 33 | 7 | R11 | 31 | 7 | R29 | 32 | 7 |
| R28 | 33 | 8 | R26 | 31 | 8 | R22 | 31 | 8 |
| R12 | 31 | 9 | R04 | 30 | 9 | R28 | 31 | 9 |
| R13 | 31 | 10 | R14 | 30 | 10 | R08 | 30 | 10 |
| R18 | 31 | 11 | R03 | 29 | 11 | R21 | 30 | 11 |
| R03 | 30 | 12 | R17 | 29 | 12 | R02 | 29 | 12 |
| R06 | 30 | 13 | R19 | 29 | 13 | R10 | 29 | 13 |
| R10 | 29 | 14 | R21 | 29 | 14 | R19 | 29 | 14 |
| R11 | 29 | 15 | R27 | 29 | 15 | R23 | 29 | 15 |
| R23 | 29 | 16 | R15 | 28 | 16 | R25 | 29 | 16 |
| R24 | 29 | 17 | R16 | 28 | 17 | R27 | 29 | 17 |
| R02 | 28 | 18 | R25 | 28 | 18 | R06 | 28 | 18 |
| R15 | 27 | 19 | R23 | 26 | 19 | R12 | 28 | 19 |

(continua)

(continuação)

| PE1 | | | PE2 | | | PE3 | | |
|---|---|---|---|---|---|---|---|---|
| Rodada | #DM | Posição | Rodada | #DM | Posição | Rodada | #DM | Posição |
| R21 | 27 | 20 | R24 | 26 | 20 | R13 | 28 | 20 |
| R25 | 27 | 21 | R02 | 25 | 21 | R24 | 28 | 21 |
| R14 | 26 | 22 | R06 | 25 | 22 | R16 | 27 | 22 |
| R04 | 25 | 23 | R12 | 25 | 23 | R17 | 27 | 23 |
| R17 | 25 | 24 | R13 | 25 | 24 | R11 | 24 | 24 |
| R27 | 25 | 25 | R22 | 21 | 25 | R15 | 24 | 25 |
| R08 | 24 | 26 | R29 | 21 | 26 | R30 | 21 | 26 |
| R19 | 24 | 27 | R10 | 19 | 27 | R26 | 20 | 27 |
| R20 | 22 | 28 | R28 | 17 | 28 | R03 | 19 | 28 |
| R01 | 21 | 29 | R05 | 13 | 29 | R09 | 16 | 29 |
| R16 | 21 | 30 | R18 | 12 | 30 | R07 | 14 | 30 |

Fonte: Dados de survey conduzido em Altamira (2005).
(1) Compreende o entorno das cidades de Altamira, Brasil Novo, Medicilândia e Uruará, no Estado do Pará.

**TABELA 3**
**Classificação da somatória de $DM\lambda_{kjl,r}$= 0 (# DM), por perfil e execução e classificação média da somatória de $DM\lambda_{kjl,r}$= 0(# DM) por execução**
**Região de estudo (1) – 2005 (n=293)**

| PE1 | | PE2 | | PE3 | | Posição |
|---|---|---|---|---|---|---|
| Rodada | Posição | Rodada | Posição | Rodada | Posição | Média |
| R01 | 29 | R01 | 2 | R01 | 3 | 11,3 |
| R02 | 18 | R02 | 21 | R02 | 12 | 17,0 |
| R03 | 12 | R03 | 11 | R03 | 28 | 17,0 |
| R04 | 23 | R04 | 9 | R04 | 4 | 12,0 |
| **R05** | **1** | **R05** | **29** | **R05** | **2** | **10,7** |
| R06 | 13 | R06 | 22 | R06 | 18 | 17,7 |
| R07 | 7 | R07 | 3 | R07 | 30 | 13,3 |
| R08 | 26 | R08 | 4 | R08 | 10 | 13,3 |
| R09 | 5 | R09 | 1 | R09 | 29 | 11,7 |
| R10 | 14 | R10 | 27 | R10 | 13 | 18,0 |
| R11 | 15 | R11 | 7 | R11 | 24 | 15,3 |
| R12 | 9 | R12 | 23 | R12 | 19 | 17,0 |
| R13 | 10 | R13 | 24 | R13 | 20 | 18,0 |
| R14 | 22 | R14 | 10 | R14 | 5 | 12,3 |
| R15 | 19 | R15 | 16 | R15 | 25 | 20,0 |
| R16 | 30 | R16 | 17 | R16 | 22 | 23,0 |
| R17 | 24 | R17 | 12 | R17 | 23 | 19,7 |
| R18 | 11 | R18 | 30 | R18 | 6 | 15,7 |
| R19 | 27 | R19 | 13 | R19 | 14 | 18,0 |
| R20 | 28 | R20 | 5 | R20 | 1 | 11,3 |
| R21 | 20 | R21 | 14 | R21 | 11 | 15,0 |
| R22 | 2 | R22 | 25 | R22 | 8 | 11,7 |
| R23 | 16 | R23 | 19 | R23 | 15 | 16,7 |
| R24 | 17 | R24 | 20 | R24 | 21 | 19,3 |
| R25 | 21 | R25 | 18 | R25 | 16 | 18,3 |
| R26 | 6 | R26 | 8 | R26 | 27 | 13,7 |
| R27 | 25 | R27 | 15 | R27 | 17 | 19,0 |
| R28 | 8 | R28 | 28 | R28 | 9 | 15,0 |
| R29 | 3 | R29 | 26 | R29 | 7 | 12,0 |

Fonte: Dados de survey conduzido em Altamira (2005).
(1) Compreende o entorno das cidades de Altamira, Brasil Novo, Medicilândia e Uruará, no Estado do Pará

A aplicação dessa rotina deve ser feita com cautela, devido ao reposicionamento dos tipos puros ao longo de execuções sucessivas – conforme já mencionado. Considerando que o mesmo tipo puro real pode mudar de posição k = n para k = m, com n ≠ m, em *R* execuções, na prática tem-se observado que a fórmula do desvio médio deve ser aplicada com a devida preocupação de reordenação dos *K* perfis, de modo que k = 1 (por exemplo) tenha estrutura semelhante de $\lambda_{kjl}$ em todas as *R* execuções. Ou seja, perfis extremos com conjuntos de probabilidades semelhantes devem estar na mesma posição e esta ordenação cabe ao pesquisador, especialmente no caso de matrizes de probabilidades aleatórias (o programa "não sabe" esta ordem). Esse procedimento evita que um falso problema de identificabilidade penalize, de forma indevida, a posição (*ranking*) daquele modelo ao longo das *R* execuções.

## Estabilidade dos parâmetros num modelo de máximo global

Uma vez solucionado o problema de *identificabilidade*, é importante observar a *estabilidade* estrutural dos parâmetros estimados pelo modelo GoM quando utilizado o programa GoM versão 3.4. Tendo em vista que o programa utiliza um algoritmo baseado em processo iterativo[6] (WOODBURY; CLIVE, 1974) para obter o valor máximo da função de máxima verossimilhança, seria esperado que os valores finais de $\lambda_{kjl}$ e $g_{ik}$ fossem estáveis. Em outras palavras, se utilizarmos os $\lambda_{kjl}$ e $g_{ik}$ estimados (após o processo iterativo com base em probabilidade designada aleatoriamente) como valores iniciais para uma nova execução do modelo (dentro do sistema de iteração dos parâmetros), seria esperado que esses parâmetros convergissem sempre para os mesmos níveis.

Na prática, no entanto, o processo iterativo do programa GoM versão 3.4[7] não estabiliza os parâmetros completamente, restando pequenas diferenças em relação aos valores estáveis finais. A ocorrência dessa instabilidade resulta do fato de a versão GoM 3.4 (e a anterior 3.3) não suprimir os valores iniciais de $\lambda_{kjl}$ estimados na primeira iteração (quando $\lambda_{kjl}$ estimados de uma rodada anterior são providos), embora o faça para os parâmetros $g_{ik}$, reduzindo a chance de replicar perfeitamente as probabilidades estimadas finais (Tabela 4).

**TABELA 4**
**Valores absolutos da média da mudança nas probabilidades estimadas por categoria de variáveis utilizadas no delineamento dos perfis extremos**
**Região de estudo (1) – 2005**

| Estatística de ajuste | Perfil extremo | R02/R01 | R03/R02 | (...) | R26/R25 | R27/R26 | R28/R27 | R29/R28 | R30/R29 |
|---|---|---|---|---|---|---|---|---|---|
| Média | PE1 | 0,3385 | 0,1599 | (...) | 0,0012 | 0,0000 | 0,0000 | 0,0000 | 0,0000 |
| | PE2 | 0,1733 | 0,1400 | (...) | 0,0288 | 0,0028 | 0,0001 | 0,0000 | 0,0000 |
| | PE3 | 0,0359 | 0,0362 | (...) | 0,0218 | 0,0000 | 0,0000 | 0,0000 | 0,0000 |
| | Média | 0,1826 | 0,1120 | (...) | 0,0173 | 0,0010 | 0,0000 | 0,0000 | 0,0000 |

Fonte: Dados de survey conduzido em Altamira (2005).
(1) Compreende o entorno das cidades de Altamira, Brasil Novo, Medicilândia e Uruará, no Estado do Pará.

[6] A partir do programa GoM 3.1 já era possível utilizar um método de solução alternativo, chamado de método gradiente. A diferença entre o método tradicional e o gradiente está no processo e não no resultado da otimização.
[7] Considerando o método padrão de entrada dos parâmetros gamma ($g_{ik}$). O método padrão assume que todos os graus de pertencimento iniciais aos perfis extremos sejam idênticos, ou seja: $g_{ik}^{inicial} = 1/k$.

Sugere-se, portanto, que, para cada execução aleatória utilizada na identificação do máximo global, sejam efetuadas *R* execuções não-aleatórias (utilizando valores de $\lambda_{kjl}$ previamente estimados) até que os valores de $\lambda_{ljk}$ se estabilizem a partir de cada execução aleatória inicial. Com os valores estabilizados, deve-se proceder a segunda execução aleatória e repetir o procedimento, até obter as 30 rodadas utilizadas na obtenção do máximo global, porém com valores dos parâmetros estabilizados nos seus valores finais.

Uma forma prática para identificar a rodada final com os parâmetros estáveis é utilizar o seguinte procedimento:

- para cada execução, *r*, com método de inclusão inicial de *lambda* aleatório (*random input lambda procedure*), tomar os valores finais de $\lambda_{kjl}$ estimados após o processo iterativo e utilizá-los como insumo inicial para uma nova execução. Dessa vez, em vez de empregar o procedimento de inclusão inicial de probabilidades aleatórias, o usuário estará partindo das probabilidades estimadas anteriormente. Esses valores utilizados, como já passaram por um processo iterativo no programa GoM versão 3.4, deveriam ser estáveis, mas não o são;
- o usuário deverá continuar utilizando sempre os valores estimados de $\lambda_{kjl}$ após a convergência da função de verossimilhança no seu valor máximo até que:

$$\sum_{L=1}^{l}\sum_{J=1}^{j}\sum_{K=1}^{k}(\lambda_{kjl,r}-\lambda_{kjl,r-1})=0 \text{ para } \forall\lambda_{kjl}$$

A execução *r - 1*, em que a condição acima for atendida, representa o modelo com estabilidade estrutural dos parâmetros finais estimados pelo GoM. A partir dessa execução, as subsequentes deverão atender à condição mencionada. Esse procedimento deve ser repetido no caso de instabilidade estrutural dos graus de pertencimento ($g_{ik}$).

## Considerações finais

O modelo de GoM, mais especificamente o *software* GoM versão 3.4, tem sido utilizado por vários pesquisadores pelas mais variadas razões, porém, há dúvidas e lacunas que permanecem quanto à implementação de todos os passos na busca do modelo mais fidedigno e, desejavelmente, único (identificável). Neste artigo, avançou-se acerca da discussão da identificabilidade do modelo *Grade of Membership* e da estabilidade dos parâmetros estimados por processo iterativo.

A importância da identificabilidade já foi discutida em trabalho anterior (CAETANO; MACHADO, 2009). Contudo, os autores não propuseram uma rotina operacional para a localização empírica de um modelo de máximo global.

Foi proposto, aqui, um procedimento simples de identificação quantitativa de um modelo de máximo global, ou seja, um modelo identificável em que seus parâmetros, $\lambda_{kjl}$ e $g_{ik}$, possuem solução única. A variabilidade presente na estatística de identificação sugerida (DM) ocorre em razão de dois possíveis fatores: erros de medição das variáveis, que podem ser repassados para o modelo final;[8] e instabilidade estrutural dos parâmetros. Assim, também sugeriu-se um procedimento que estabilize os parâmetros obtidos por meio de processo iterativo.

O critério utilizado para identificação do máximo global pode ser influenciado pelo

[8] Erros na medição de uma variável em particular podem gerar imprecisões nas associações que se desejam estudar. Por exemplo, utilizar uma variável como autopercepção de saúde para relacionar com a ocorrência de uma determinada doença pode levar a conclusões espúrias, devido aos vieses já estabelecidos na medição desta variável (SILVEIRA et al., 2002). O método GoM, que atua por meio da geração de novas variáveis contínuas com base nas associações presentes entre um grupo de variáveis, ameniza este problema, gerando uma variável latente, por exemplo, "saúde", que agrega características de todas as variáveis relacionadas ao conceito de saúde.

número de execuções (r). Possivelmente, quanto maior o número de execuções, mais certamente chegar-se-á próximo do máximo global. Experiências com diferentes bases de dados dos autores deste estudo indicam que, em simulações com 20, 30, 50 e 100 execuções aleatórias iniciais, o máximo global tendia a se repetir a partir de 30 execuções, consideradas o número mínimo necessário de execuções.

As vantagens e desvantagens relativas ao uso do *software* GoM 3.4 já foram descritas em trabalhos prévios (ver, por exemplo, PEREIRA et. al., 2007). Os procedimentos aqui sugeridos pretendem tornar a utilização empírica do modelo *Grade of Membership* mais intuitiva para os usuários finais. Tendo em vista a necessidade de uma matriz inicial para que o processo de convergência ocorra, procurou-se explicitar claramente o que fazer para se obter um modelo final que seja adequado, estabelecendo um conjunto de etapas a serem seguidas. Além disso, é importante que os pesquisadores tomem conhecimento de como responder a críticas em relação à identificabilidade, uma potencial fonte de resistência à aplicação de modelagem baseada no GoM. A aplicação empírica do GoM na área das ciências sociais tem-se resumido à construção de perfis e análise de prevalências. Alguns estudos recentes avançaram utilizando o método para definição de hierarquias (GARCIA et al., 2007; GUEDES et al., 2009a, b e d). O modelo GoM, no entanto, ainda possui potenciais inexplorados nas áreas de ciências sociais aplicadas, como a análise de prognósticos (MANTON et al., 1994).

Para as ciências sociais, o método GoM é uma ferramenta estratégica, uma vez que é da natureza desse campo do conhecimento trabalhar com variáveis categóricas e/ou com a categorização das variáveis contínuas para análise comparativa. A maioria dos métodos de análise multivariada demanda variáveis contínuas, o que torna o método GoM atrativo. Ademais, o GoM pode ter outros fins além da identificação de padrões de associação que tipifiquem os elementos de um conjunto de forma mais condizente com a complexidade da realidade social, na qual a dicotomia pertencimento-não pertencimento a um determinado conjunto de características específicas raramente é válida. Os perfis delineados podem ser utilizados, por exemplo, para estabelecer os critérios para o recrutamento da participação em grupos focais ou entrevistas em profundidade (MIRANDA-RIBEIRO et al., 2007) e para a identificação de variáveis relevantes na composição de indicadores sintéticos. Essas qualidades fazem do GoM uma alternativa viável para análise de bancos de dados de complexidade variável, além de oferecer suporte para estudos interdisciplinares colaborativos que incluem abordagens quantitativas e qualitativas.

Cabe observar que os procedimentos aqui propostos constituem um primeiro passo em direção à busca de um modelo estável. É necessário que os usuários conheçam as potencialidades, limitações e procedimentos necessários para a construção de um modelo que descreva as associações implícitas e revele os padrões mais frequentes dos dados de forma fidedigna. Trabalhos futuros poderiam enfatizar a possibilidade de se gerar uma estimativa de intervalo de confiança para a identificabilidade. Uma ideia seria que, conhecendo os valores de $\lambda_{jl}$ para cada categoria em várias execuções ($r$) para um mesmo perfil ($k$), basta que se calcule o erro-padrão de DM e se estime o intervalo de confiança. Assim, pode-se obter o máximo global baseando-se nos valores inferior e superior do intervalo de confiança a 5% de significância.

Finalmente, é importante observar que este trabalho possui uma limitação. Os procedimentos técnicos sugeridos foram aplicados a apenas uma base de dados e seria de grande utilidade empregar estas técnicas a outras bases empíricas. Com efeito, existem situações nas quais, naturalmente, não seria possível encontrar grupos ou perfis, no caso de bancos de dados com entropia máxima, em que cada indivíduo na amostra seria tão diferente que a convergência em torno de um conjunto de probabilidades estimadas definidoras de tipos puros não poder-se-ia concretizar (PEREIRA et al., 2007). Nestes casos, a procura de agrupamentos não seria uma estratégia de análise adequada.

**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.

CASSADY, F.; PIEPER, C. F.; CARROL, B. J. Subtypes of mania determined by Grade of Membership Analysis. **Neuropsychopharmacology**, v. 25, n. 3, p. 373-383, 2001.

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.

DECISION SYSTEMS INC, s.d. Disponível em: <http://www.dsisoft.com/grade_of_membership.html>. Acesso em: 15 mar. 2009.

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.; COSTA, S. M.; BRONDIZIO, E. S. Revisiting the hierarchy of urban areas in the brazilian Amazon: a multilevel approach. **Population & Environment**, v. 30, p. 159-192, 2009a.

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, 2009b.

GUEDES, G. R.; QUEIROZ, B. L.; VANWEY, L. K. Transferências intergeracionais privadas na Amazônia rural brasileira. **Nova Economia**, v. 19, n.2, 2009c.

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, 2009d.

GUIMARÃES, M. D. C.; OLIVEIRA, H. N.; CAMPOS, L. N.; SANTOS, C. A.; GOMES, C. E. R.; OLIVEIRA, S. B.; FREITAS, M. I. F.; ACURCIO, F. A.; MACHADO, C. J. Reliability and validity of a questionnaire on vulnerability to sexually transmitted infections among adults with chronic mental illness: PESSOAS Project. **Revista Brasileira de Psiquiatria**, v. 30, n. 1, p. 55-59, 2008.

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.

PEREIRA, C. C. A.; MACHADO, C. J.; RODRIGUES, R. N. Perfis de causas múltiplas de morte relacionadas ao HIV/AIDS nos municípios de São Paulo e Santos, Brasil, 2001. **Cadernos de Saúde Pública**, v. 23, n. 3, p. 645-655, 2007.

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.

SILVEIRA, M. F.; BERIA, J.; HORTA, B. L.; TOMASI, E. Self-assessment of STD/AIDS vulnerability among women, Brazil. **Rev. Saúde Pública**, v. 36, n. 6, p. 670-677, 2002.

VANWEY, L. K.; GUEDES, G. R.; D'ANTONA, A. O. Land use trajectories after migration and land turnover. In: POPULATION ASSOCIATION OF AMERICA ANNUAL MEETING. **Anais…** New Orleans, 2008.

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**

*Identificabilidad y estabilidad de los parámetros en el método Grade of Membership (GoM): Consideraciones metodológicas y prácticas*

El método Grade of Membership (GoM) ha sido cada vez más utilizado por los demógrafos brasileños y tiene la ventaja de poseer un parámetro que mide la heterogeneidad individual, sobre la base de las correlaciones no observables entre las categorías de respuesta de las variables de interés, generando una medida del grado de pertenencia de cada individuo a perfiles extremos. Algunos autores, sin embargo, destacan cuestiones importantes en la calibración de los modelos finales que utiliza el programa GoM versión 3.4, como el problema de identificabilidad – soluciones múltiples para parámetros estimados. En este artículo, se sugiere un procedimiento capaz de identificar un modelo final con una solución única que describa los tipos puros de mayor fidelidad con respecto a la base de datos, con una intención de optimización. Para ilustrar este proceso, se utilizó una base de dados correspondiente a un relevamiento económico y socio-demográfico de una población de pequeños agricultores residentes a lo largo de la Autopista Transamazônica, en el Estado de Pará. También se identificó la existencia de inestabilidad en los parámetros estimados por el programa GoM 3.4, y se propuso un método de estabilización de sus valores. Con esos procedimientos combinados, los usuarios del programa GoM 3.4 podrán describir su base de dados en forma más adecuada y responder a las críticas sobre cuestiones de identificabilidad y estabilidad de los modelos resultantes. Estas soluciones empíricas son relevantes porque afectan cálculos de superioridad y de incidencia de eventos de interés, además de traer consecuencias importantes sobre el punto y el momento correctos para las intervenciones de políticas públicas o de planificación prospectiva en análisis de proyección.

**Palabras-clave**: Grade of Membership. Identificabilidad. Estabilidad. Máximo global. Conjuntos nebulosos.

**Abstract**

*Identifiability and stability of standards in the Grade of Membership (GoM) method: methodological and practical considerations*

The Grade of Membership (GoM) method has been increasingly employed by Brazilian demographers, and has the advantage of including a parameter that measures individual heterogeneousness on the basis of non-observable correlations among the categories of

responses to variables of interest. The parameter shows each individual's degree of membership to extreme profiles. Several authors, however, have called attention to important issues in adjusting the final models that use 3.4 Version of the GoM Program, such as the problem of identifiability – multiple solutions for estimated parameters. In this article a procedure is discussed that is able to identify a final model with a single solution that describes the pure types that are the most reliable for the database, in an attempt at streamlining. To illustrate this process, a database was used with data corresponding to an economic and sociodemographic study of a population of small farmers living along the TransAmazon Highway, in the northern State of Pará, Brazil. The existence of instability in the parameters estimated by the GoM 3.4 Program was also identified and a method of stabilization of its values was proposed. With these combined procedures, users of the GoM 3.4 Program will be able to describe their databases more adequately and respond to criticisms regarding the identifiability and stability of the resulting models. These empirical solutions are significant. Not only do they affect calculations of prevalence and incidence of events of interest, they also bring about important consequences at the correct point and correct moment for interventions of public policies or of prospective planning in projection analyses.

Keywords: Grade of Membership Method. Identifiability. Stability. Global maximum. Fuzzy sets.

Recebido para publicação em 23/12/2009
Aceito para publicação em 31/03/2010