Aspectos práticos na identificação de um modelo *Grade of Membership* (GoM) de máximo global: o uso da moda das probabilidades estimadas

Gilvan Ramalho Guedes*
Pamila Cristina Lima Siviero**
Carla Jorge Machado***

Introdução

O método *Grade of Membership* (GoM) é um instrumento de análise reconhecidamente útil na descrição de bases de dados complexas, com caráter multidimensional. Possui a vantagem de não demandar grandes tamanhos de amostra e poder ser utilizado com variáveis endógenas (MANTON; WOODBURY; TOLLEY, 1994). O programa GoM 3.4, que estima funções de pertinência derivadas do algoritmo GoM proposto por Woodbury e Clive (1974), vem sendo amplamente empregado por pesquisadores brasileiros. Estudos anteriores (GUEDES et al., 2010a e 2010b; CAETANO; MACHADO, 2009) discutiram a ideia de que os modelos gerados pelo programa GoM 3.4, partindo de uma matriz aleatória inicial de probabilidades, devem ser cuidadosamente comparados para que, ao longo de sucessivas execuções, o pesquisador possa se assegurar de escolher o modelo que se refere ao padrão mais recorrente das probabilidades finais obtidas. Nesse contexto, no qual o processo de obtenção de matrizes de probabilidades finais é iterativo, necessita-se ter cautela sobre a correta identificação do modelo final elegido, uma vez que se pode obter, em uma determinada execução, um modelo de máximo local e não de máximo global.

É essa ideia de se obter um padrão recorrente que os autores têm chamado de processo de localização de um modelo de máximo global. A necessidade de identificar um modelo final de máximo global ocorre porque o programa GoM 3.4 não assegura que, dada qualquer matriz inicial aleatória das probabilidades de pertencimento de cada categoria das variáveis aos perfis extremos definidos, os resultados serão sempre os mesmos. Contudo, ao longo de sucessivas execuções, um padrão de probabilidades emerge e é entre esse conjunto de modelos com padrões similares de probabilidades que se encontra o modelo com o valor máximo da função de verossimilhança (CAETANO; MACHADO, 2009).

Apesar de iniciativas anteriores para localizar o modelo que representa o padrão mais recorrente dos parâmetros finais estimados, as estatísticas de localização propostas não consideram completamente a ideia de "padrão mais frequente", pois utilizam a média e não a moda das probabilidades estimadas. Como a média não é necessariamente representativa de qualquer valor na população, é possível, até mesmo provável, que a probabilidade de cada uma das *r* execuções não coincida com a média das R execuções. Como a moda apreende o conceito de "valor mais frequente", nessa nota metodológica propõe-se readaptar a medida de localização de máximo global sugerida por Guedes et al. (2010 e 2011), substituindo a média pela moda como me-

* 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 Population Dynamics, professora adjunta nível III do Centro de Desenvolvimento e Planejamento Regional – Cedeplar/UFMG.

dida de referência, na estatística de desvio em relação a cada probabilidade estimada.

Antecedentes metodológicos

Os modelos empíricos GoM[1] baseados em matrizes de probabilidades iniciais aleatórias ('parâmetros λ_{kjl}'[2] iniciais necessários para dar início ao processo iterativo) podem diferir entre uma execução e outra (CAETANO; MACHADO, 2009). Nesse sentido, os resultados finais não seriam identificáveis, já que uma mesma categoria de resposta de uma variável específica poderia ter mais de uma probabilidade de ocorrência entre modelos empíricos distintos.

A mudança nas probabilidades finais estimadas entre execuções sucessivas baseadas em parâmetros iniciais aleatórios não é uma falha do método em si, mas antes uma característica inerente aos métodos de conglomerados que se baseiam na distância de cada elemento aos centroides dos grupos. Ou seja, cada vez que o centroide de um grupo é alterado, a distância de cada elemento ao centroide é também modificada, resultando em grupos finais com elementos diferentes a cada modificação inicial (GAN; MA; WU, 2007).

A ausência de um critério de identificabilidade para modelos empíricos derivados do método GoM foi e ainda é uma fonte de críticas aos trabalhos que utilizam o modelo, embora muitas das críticas sejam baseadas em fundamentos não diretamente aplicados à lógica de agrupamento, mas sim em relação a métodos de regressão. Com efeito, a questão da identificabilidade é um ponto crucial na análise de dados, pois, desse modo, parâmetros estimados podem ser replicados.

Apoiando-se na ideia proposta por Caetano e Machado (2009), Guedes et al. (2010) sugeriram que um modelo empírico identificável (isto é, aquele com parâmetros únicos) poderia ser aproximado por meio da utilização de uma medida denominada *Desvio em Relação à Média das Probabilidades Finais* ($DM_{kjl,r}$). Segundo os autores, em sucessivas execuções, é possível aproximar um modelo identificável com base na contagem do número de probabilidades de ocorrência das categorias (l) das variáveis internas (j) empregadas para definir cada perfil extremo (k) que apresenta desvios-nulo em relação à média das probabilidades estimadas para a mesma categoria ao longo de R execuções. Quanto maior o número de desvios-nulo em relação à média, mais próxima uma execução estaria de um modelo de solução única.

Em artigo posterior (GUEDES et al., 2011), respondendo a críticas e sugestões de revisores, os autores propuseram um novo localizador de um modelo estável (chamado modelo de máximo global), incorporando a contribuição dos desvios-padrão dos $DM_{kjl,r}$ ao longo das *l* categorias. A nova medida de localização, chamada de localizador MGP (Máximo Global Ponderado), é obtida a partir da divisão do desvio-padrão dos $DM_{kjl,r}$ pelo número de vezes que $DM_{kjl,r} = 0$. O localizador MGP, nesse sentido, aproxima-se da ideia de um coeficiente de variação, pois tenta incorporar a contribuição da dispersão dos $DM_{kjl,r}$ sobre a representatividade da contagem de desvios-nulo em cada uma das execuções (ou modelos).

O localizador MGP representou um avanço por criar uma regra de decisão aos usuários do programa GoM 3.4 entre, por exemplo, dois modelos que apresentem o mesmo número de $DM_{kjl,r} = 0$, porém com desvios-padrão de $DM_{kjl,r}$ distintos. Assim, quanto maior for o desvio-padrão dos $DM_{kjl,r}$, independentemente do número de desvios-médio nulos, menor será a chance de aquela execução representar um modelo de solução única. Ao mesmo tempo, entre modelos com desvios-padrão idênticos, porém com número de desvios-nulo distintos, seria preferido o modelo com maior número de desvios-nulo, remetendo ao localizador

[1] Estamos chamando "modelos empíricos GoM" aqueles resultantes do programa GoM 3.4.

[2] O parâmetro lambda é chamado de parâmetro locacional na literatura técnica sobre *Grade of Membership* porque representa a probabilidade de ocorrência da categoria *l* de uma variável *j* em um perfil *k*. Assim, essa probabilidade ajuda a localizar o vértice K=k num simplex (politopo) de dimensão *K* (MANTON; WOODBURY; CLIVE, 1994).

originalmente proposto no primeiro artigo dos autores (GUEDES et al., 2010).

Redefinindo o critério de identificação de um modelo estável

Apesar do importante avanço para identificar um modelo estável, o localizador MGP apresenta uma falha potencial. A ideia na qual ele está baseado é de que existe um padrão recorrente que emerge ao longo de sucessivas execuções. No entanto, o localizador baseia-se numa estatística de desvio de cada probabilidade (numa execução *r*) em relação à média dessas probabilidades (em R execuções). Para representar com maior propriedade a ideia de valor recorrente, a melhor medida-resumo poderia não ser a média, mas sim a moda das probabilidades estimadas. Por definição, a moda representa o valor mais frequente numa série de elementos; a média, por outro lado, pode não representar nenhum dos valores da série, especialmente quando a distribuição dos elementos apresenta elevada dispersão. Além de ser mais coerente com a noção inicial de modelo estável baseado em valores mais frequentes, a moda evita potenciais casos de inexistência de valores de probabilidades idênticas à média dos lambdas.

Assim, propõe-se um *desvio em relação à moda*, que pode ser obtido da seguinte forma:

$$DMOD_{kjl,r} = \lambda^{r}_{kjl} - MOD^{R}_{r=1}(\lambda^{r}_{kjl})$$

onde:

λ^{r}_{kjl} é a probabilidade estimada de ocorrência de resposta da categoria *l* da variável *j* no perfil *k* para a execução *r*; e $MOD^{R}_{r=1}(\lambda^{r}_{kjl})$ corresponde à moda das probabilidades estimadas da mesma categoria *l* ao longo das R execuções.

A partir de $\mathrm{DMOD}_{\mathrm{kjl,r}}$, o cálculo do localizador MGP segue a mesma lógica proposta por Guedes et al. (2011). Ou seja:

$$MGP = \frac{\sigma^{L}_{l=1}\left(DMOD_{kjl,r}\right)}{\sum_{l=1}^{L} \#\, DMOD_{kjl,r} = 0}$$

onde:

$\sigma^{L}_{l=1}\left(DMOD_{kjl,r}\right)$ representa o desvio-padrão dos $\mathrm{DMOD}_{\mathrm{kjl,r}}$ ao longo das L categorias de cada execução *r*.

Apesar de a aplicação da estatística $\mathrm{DMOD}_{\mathrm{kjl,r}}$ parecer direta, ocorrerão casos em que, mesmo em sucessivas execuções, não existirá uma moda para algumas das categorias das variáveis incluídas no modelo. Ainda em relação à mesma variável, pode ocorrer de algumas de suas categorias apresentarem probabilidades estimadas repetidas e outras não (caso comum em probabilidades extremas – por exemplo, λ_{kjl} = 0,0000 ou λ_{kjl} = 1,0000). Para esses casos, sugerimos a utilização do valor médio das probabilidades das R execuções para a categoria específica, de modo que essas categorias também possam entrar no cálculo do desvio-padrão, utilizado na obtenção do localizador MGP.

O procedimento sugerido neste trabalho apenas substitui o desvio em relação à média pelo desvio em relação à moda das probabilidades. Essa alteração não modifica algumas limitações desse tipo de procedimento pós-estimação, que é a sua dependência assintótica ao número de modelos gerados. Ou seja, o localizador MGP baseado no $\mathrm{DMOD}_{\mathrm{kjl,r}}$, assim como o MGP baseado no $\mathrm{DMOD}_{\mathrm{kjl,r}}$, é tão mais preciso quanto maior for o número de execuções (R). Entretanto, regra geral, 30 execuções são suficientes, pois o valor da função de verossimilhança não é substantivamente alterado com o aumento das execuções. Contudo, mesmo tendo esta regra norteadora, é sempre importante ter em mente que depende do pesquisador perceber se, em determinadas circunstâncias, é necessário obter um número maior de execuções.[3]

[3] Uma forma de testar essa proposição é obter dez execuções adicionais e efetuar o teste da razão de verossimilhança. O teste está disponível em Manton, Woodburry e Tolley (1994) e deve comparar um modelo identificado com 30 execuções aleatórias e outro modelo identificado com 40 execuções aleatórias. Um resultado significativo para o teste é suficiente para impor um limite ao número de execuções adicionais e optar pela linha de base (30 execuções).

Comparando os indicadores $DM_{kjl,r}$ e $DMOD_{kjl,r}$

Utilizou-se uma amostra sobre sistemas de uso do solo de pequenos agricultores da região da Rodovia Transamazônica, no Pará, de 2005, para ilustrar a utilização da estatística $DM_{kjl,r}$ e $DMOD_{kjl,r}$. Essa é a mesma base de dados empregada por Guedes et al. (2010 e 2011). A proposta consiste em ilustrar a substituição da média pela moda na estatística de máximo global ponderado (MGP). Apresenta-se, aqui, a comparação apenas para um perfil extremo.

Cada execução aleatória do programa retorna uma probabilidade estimada para cada perfil k, em cada variável j, em cada nível de resposta l. Esse valor é representado por λ_{kjl}. Como o modelo é iterativo e, em cada execução, a matriz inicial de probabilidades é aleatória, espera-se que o valor de λ_{kjl} varie de uma execução para outra.

No primeiro caso calculou-se a estatística de MGP utilizando o desvio em relação à média dos λ_{kjl} em sucessivas execuções. Primeiramente, executou-se o programa 30 vezes e foram organizadas as colunas referentes ao perfil extremo ($k=1$, por exemplo) em uma mesma planilha de dados, obtendo, a partir daí, a média dos λ_{kjl} na linha. Em outras palavras, a média é calculada com base nas probabilidades estimadas para uma mesma variável j, em uma mesma categoria l, em execuções *distintas.* Uma vez calculada a média, o passo seguinte consiste em calcular o desvio de cada uma dessas probabilidades estimadas em relação à sua média, o desvio-médio (DM_{kjl}), que consiste em subtrair de cada λ_{kjl} o $\lambda_{kjl\ médio}$.

A primeira parte da Tabela 1 apresenta os valores dos DM_{kjl} para três variáveis selecionadas do perfil extremo $k=1$. Cada coluna da Tabela representa uma execução (RA01 a RA10) e cada célula indica o quanto cada um dos λ_{kjl} desviou de seu valor médio ($\lambda_{kjl\ médio}$). Na sequência, contabilizou-se, em cada execução (em cada uma das colunas), o número de vezes que o DM_{kjl} foi igual a zero. O passo seguinte consiste em calcular o desvio-padrão dos DM_{kjl}, para cada uma das *execuções*, ou seja, em cada uma das colunas. Diferentemente da média, o desvio-padrão dos DM_{kjl} é calculado com base nas probabilidades estimadas para uma *mesma* execução, com todas as l categorias de todas as j variáveis. Enquanto a média é calculada na linha (mesma variável j, mesma categoria l, execuções distintas), o desvio-padrão é calculado na coluna (*mesma* execução, todas as j variáveis com suas l categorias *distintas*). Assim, o MGP calculado com base no DM_{kjl} pode ser calculado dividindo-se o desvio-padrão dos DM_{kjl} pelo número de DM_{kjl} iguais a zero.

Para a moda, o processo de cálculo é semelhante. No entanto, na etapa na qual foi calculada a média de cada um dos λ_{kjl}, basta substituir esse valor médio pelo valor *modal,* caso esta moda exista*.* Se não existir, mantêm-se os desvios em relação à média. As demais etapas são idênticas às descritas anteriormente. A segunda parte da Tabela 1 apresenta, para as mesmas três variáveis e execuções aleatórias, os valores dos desvios em relação à *moda* ($DMOD_{kjl}$).

É possível perceber, na Tabela 1, o efeito da substituição do desvio em relação à *média* (DM_{kjl}) pelo desvio em relação à *moda* ($DMOD_{kjl}$). Em situações nas quais não havia moda (variável 1, categorias 1 e 2; variável 2, categoria 1; variável 3, categorias 0 a 2), manteve-se o desvio em relação à média. Observando o primeiro quadrante da Tabela 1, se o modelo consistisse apenas dessas três variáveis, nas dez execuções aleatórias teríamos o mesmo número de DM_{kjl} (1). Nesse sentido, o MGP teria apenas um parâmetro – o desvio-padrão dos DM_{kjl} – e o melhor modelo seria aquele com menor desvio-padrão. No caso da moda, a situação é distinta. O número de desvios em relação à moda iguais a zero variou entre as execuções e a maior contagem foi observada apenas nas execuções aleatórias 4 e 8 (segunda parte da Tabela 1). Nesse caso, ambos os parâmetros contribuíram para a escolha do melhor modelo, aquele que mais se aproxima do padrão frequente que emerge dos dados, e o menor desvio-padrão definiu o melhor modelo entre as execuções 4 e 8 (Tabela 1).

TABELA 1
Valores de $DM_{kjl,r}$ e $DMOD_{kjl,r}$ derivados das probabilidades estimadas de modelos GoM (perfil k=1 de K=3) em dez execuções utilizando variáveis selecionadas para fins ilustrativos
Área de estudo de Altamira, Pará – 2005

Variável (j)		Execução (r)									
Rótulo	Categoria (l)	RA01	RA02	RA03	RA04	RA05	RA06	RA07	RA08	RA09	RA10
		$DM_{kjl,r}$									
var1	0	-0,0306	-0,0306	-0,0306	-0,0306	0,0965	-0,0306	-0,0039	-0,0306	0,0907	-0,0306
	1	0,0131	-0,0041	0,0093	-0,0247	0,0292	-0,0013	-0,0259	0,0084	-0,0036	-0,0153
	2	0,0175	0,0347	0,0213	0,0553	-0,1257	0,0319	0,0298	0,0222	-0,0871	0,0459
	3	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
var2	0	-0,1476	-0,1476	-0,1476	-0,1476	0,4307	-0,1476	0,1444	-0,1476	0,3101	-0,1476
	1	0,1333	0,0995	0,0852	0,1344	-0,3293	0,1036	-0,1049	0,1132	-0,2353	0,0939
	2	0,0283	0,0561	0,0457	0,0272	-0,0874	0,0567	-0,0874	0,0484	-0,0874	0,0401
	3	-0,0140	-0,0080	0,0166	-0,0140	-0,0140	-0,0127	0,0478	-0,0140	0,0126	0,0136
var3	0	-0,0962	-0,1493	0,0126	-0,0761	0,1293	-0,0487	0,1181	-0,0163	0,1270	-0,0050
	1	0,0318	0,0587	-0,0009	0,0475	-0,0649	0,0335	-0,0580	0,0115	-0,0590	0,0041
	2	0,0257	0,0426	0,0008	0,0412	-0,0519	0,0277	-0,0476	0,0173	-0,0554	0,0134
	3	0,0387	0,0480	-0,0124	-0,0124	-0,0124	-0,0124	-0,0124	-0,0124	-0,0124	-0,0124
$\#DM_{kjl,r} = 0$		1	1	1	1	1	1	1	1	1	1
		$DMOD_{kjl,r}$									
var1	0	0,0000	0,0000	0,0000	0,0000	0,1271	0,0000	0,0267	0,0000	0,1213	0,0000
	1	0,0131	-0,0041	0,0093	-0,0247	0,0292	-0,0013	-0,0259	0,0084	-0,0036	-0,0153
	2	0,0175	0,0347	0,0213	0,0553	-0,1257	0,0319	0,0298	0,0222	-0,0871	0,0459
	3	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
var2	0	0,0000	0,0000	0,0000	0,0000	0,5783	0,0000	0,2920	0,0000	0,4577	0,0000
	1	0,1333	0,0995	0,0852	0,1344	-0,3293	0,1036	-0,1049	0,1132	-0,2353	0,0939
	2	0,1157	0,1435	0,1331	0,1146	0,0000	0,1441	0,0000	0,1358	0,0000	0,1275
	3	0,0000	0,0060	0,0306	0,0000	0,0000	0,0013	0,0618	0,0000	0,0266	0,0276
var3	0	-0,0962	-0,1493	0,0126	-0,0761	0,1293	-0,0487	0,1181	-0,0163	0,1270	-0,0050
	1	0,0318	0,0587	-0,0009	0,0475	-0,0649	0,0335	-0,0580	0,0115	-0,0590	0,0041
	2	0,0257	0,0426	0,0008	0,0412	-0,0519	0,0277	-0,0476	0,0173	-0,0554	0,0134
	3	0,0511	0,0604	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
$\# DMOD_{kjl,r} = 0$		4	3	4	5	4	4	3	5	3	4

Fonte: Base de dados de Altamira, 2005. ACT (2010),
Nota: RA = Rodada com matriz de lambdas (λ_{kjl}) iniciais aleatórios.

Conclusão

Com a finalidade de facilitar e estimular o uso do programa GoM 3.4 para identificação de perfis subjacentes aos dados que se deseja trabalhar, este texto avança ao propor uma nova medida de localizador de um modelo de máximo global, que leva em consideração a moda das probabilidades estimadas. Retomando a ideia já explicitada, se é desejado o valor mais frequente, a melhor medida-resumo é a moda. Contudo, este indicador não é isento de limitações, sendo necessária cautela para o seu uso. A moda pode se repetir pouquíssimas vezes em uma série de valores, não se revelando representativa.

Os próximos passos deste trabalho constituem ir além do processo de identificação de máximo global, procurando abarcar, em artigos futuros, o uso do método, com o auxílio do programa GoM 3.4, em situações em que a hierarquia (ou a gradação), tal como estado de saúde "muito bom" ao "muito grave" entre sucessivos perfis, é um traço relevante a ser demarcado. Considera-se importante, assim, incorporar, explicitamente, o contexto dentro do qual o trabalho é desenvolvido.

Referências

ACT – Anthropological Center for Training and Research on Global Environmental Change. Base de dados da região de estudo de Altamira, Pará – Microdados da pesquisa de 2005. Bloomington, USA, 2010 (Dados não publicados).

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.

GAN, G.; MA, C.; WU, J. **Data clustering**: theory, algorithms, and applications. SIAM, Philadelphia, 2007.

GUEDES, G. R.; MACHADO, A. C.; MACHADO, C. J.; BRONDÍZIO, 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.; SIVIERO, P. C. L.; CAETANO, A. J.; MACHADO, C. J.; BRONDÍZIO, E. S. Incorporando a variabilidade no processo de identificação do modelo de máximo global no Grade of Membership (GoM): considerações metodológicas. **Revista Brasileira de Estudos de População**, v. 28, n. 2, 2011 (no prelo).

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

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.

Recebido para publicação em 08/06/2011
Aceito para publicação em 30/08/2011