O ambiente escolar no desempenho acadêmico do aluno: criação de uma escala a partir do SAEB-99*

KAIZÔ IWAKAMI BELTRÃO
Escola Nacional de Ciências Estatísticas (ENCE/IBGE)
kaizo@ibge.gov.br

IURI COSTA LEITE
Escola Nacional de Saúde Pública (ENSP/FIOCRUZ)
iuri@procc.fiocruz.br

MARIA EUGÊNIA FERRÃO
Escola Nacional de Ciências Estatísticas (ENCE/IBGE)
mariabarbosa@ibge.gov.br

Resumo
Este artigo é parte de uma pesquisa mais ampla que investiga os fatores que contribuem para o efeito escola no desempenho de estudantes brasileiros, utilizando os dados do Sistema de Avaliação da Educação Básica (SAEB), complementados pelos do censo escolar. Um dos fatores investigados é o ambiente escolar, cujas informações foram levantadas no questionário de escolas da pesquisa. Foi construída uma escala através da técnica GOM (Grade of Membership) de conjuntos difusos. O objeto deste artigo é a apresentação e a aplicação dessa técnica aos dados do SAEB referentes às escolas. A escala resume características da escola, tais como nível de ruído na sala de aula, grau de limpeza do ambiente, disponibilidade de banheiros, biblioteca e laboratórios, segurança da escola, etc. Essa técnica permite uma redução da dimensionalidade do questionário e a sua subseqüente utilização como covariável de um modelo de regressão.
Palavras chave: conjuntos difusos, redução de dimensionalidade, escalas.

Resumen
Este artículo es parte de una investigación más amplia sobre los factores que contribuyen al efecto escuela en el desempeño de los estudiantes brasileños, utilizando los datos del Sistema de Evaluación de la Educación Básica (SAEB) complementados por los datos del censo escolar. Uno de los factores investigados es el ambiente de la escuela, con informaciones que han sido levantadas en uno de los cuestionarios del SAEB. Una escala ha sido construida y se utilizó la técnica GOM (Grade of Membership) de conjuntos difusos. El objeto de este artículo es la presentación y aplicación de esa técnica a los datos del SAEB respecto a las escuelas. La escala resume las características de la escuela como, por ejemplo, el nivel de ruido en las aulas, el grado de limpieza del ambiente, disponibilidad de lavabos, biblioteca y laboratorios, seguridad en la escuela y otros factores. Esa técnica permite una reducción de la dimensionalidad del cuestionario y su siguiente utilización como co-variable de un modelo de regresión.
Palabras-clave: conjuntos difusos, reducción de dimensionalidad, escalas.

* Trabalho apresentado originalmente em Encontro Regional da ABE em Fortaleza, 2002.

Abstract
This article is part of an ongoing research project on factors explaining school-effect on students' proficiency. The research uses Elementary Education Evaluation System (SAEB) data complemented by School Census information. Both surveys are conducted by INEP, an institute linked to the Ministry of Education. One of the factors under scrutiny is school environment and infrastructure, data collected in one of the SAEB questionnaires. A scale was drawn up using the fuzzy sets GOM – Grade of Membership – technique. The goal of this article is to present the technique and then apply it to the SAEB data on schools. The scale summarizes characteristics of the school, such as noise level in classrooms, tidiness, availability of restrooms, libraries and labs, as well as school safety and other factors. This technique allows for a reduction in the dimensionality of the questionnaire for its subsequent use as a covariate in regression models, and specifically in our research project, in a multilevel regression.
Key words: fuzzy sets, reduction of dimensionality, scales.

1. Introdução

Um problema recorrente na análise de dados é a necessidade de se reduzir a dimensão de um conjunto de variáveis, seja para reconhecer grupos específicos, seja para uma futura utilização em regressões. Uma técnica eficiente, porém pouco explorada em dados nacionais, é o "grade of membership" (GOM, grau de pertinência). Esta técnica tem como embasamento a teoria de conjuntos difusos (Zadeh, 1965; Woodbury, Clive, 1974). Uma vantagem de tal técnica é a quantificação de possíveis ambigüidades sobre a alocação de elementos a conjuntos. Isto é importante, pois na maior parte dos casos encontrados na prática os conjuntos não têm contornos bem definidos, e elementos individuais podem apresentar características de mais de um dos conjuntos. Podemos citar ainda outras duas vantagens importantes desta técnica: na sua implementação, não se faz necessário o conhecimento prévio dos conjuntos nos quais queremos classificar os pontos; e vetores com alguns dados omissos não precisam ser descartados da análise; situações com as quais os pesquisadores se deparam constantemente.

O objetivo deste artigo é apresentar e discutir os resultados da aplicação dessa técnica a um conjunto de dados utilizados em um estudo dos fatores associados à proficiência do alunado do ensino básico brasileiro (Barbosa et al. 2000 e 2001). A idéia subjacente à análise é a de que, ao se investigar o efeito do ambiente escolar (tome-se este como sendo representado pelas condições de funcionamento das salas de aula, além da infra-estrutura física e de equipamento) no desempenho acadêmico do aluno, ao invés de utilizarmos todas as variáveis originais do questionário da escola, possamos usar apenas o escore extraído a partir delas através da técnica de GOM. No estudo original, características do aluno, dos professores e da escola foram investigadas. Entretanto, vamos nos ater, aqui, somente às informações relacionadas às características físicas das escolas.

Este relatório está organizado como se segue: na segunda seção, descrevem-se o problema específico que motivou a aplicação do método, bem como as bases de dados utilizadas; na terceira seção, a metodologia é apresentada; a quarta seção compreende os resultados obtidos. Finalmente, na última seção, apresenta-se uma discussão sucinta dos achados, destacando-se possíveis desdobramentos deste estudo.

2. Base de dados

O estudo motivador foi o ajuste de um modelo multinível aos dados do SAEB-99 (Barbosa et al., 2001), que procurou mensurar o efeito da escola na proficiência dos alunos. Nesse estudo, utilizam-se as informações do Sistema de Avaliação do Ensino Básico (SAEB) na sua edição de 1999. O SAEB é um levantamento em ampla escala realizado pelo Instituto Nacional de Estudos e Pesquisas Educacionais (INEP) do Ministério da Educação. Esse levantamento tem representatividade nacional e estadual, com base em uma amostragem de alunos das 4ª e 8ª séries do primeiro grau e da 3ª série do segundo grau. O objetivo principal do SAEB é medir a proficiência dos alunos em conhecimentos básicos. Além do questionário de conteúdo aplicado aos alunos, são também coletadas informações sobre a escola, diretores e professores. As variáveis utilizadas, em nosso estudo, referem-se às características de infra-estrutura física das escolas. Além das informações contidas no questionário do SAEB para escolas, acrescentaram-se três variáveis retiradas do censo escolar, a saber: a existência ou não de biblioteca, de laboratório de ciências e de laboratório de informática. A lista de variáveis encontra-se na tabela 1.

Tabela 1 – Variáveis utilizadas na análise

VARIÁVEL	DESCRIÇÃO
E_1_1X	Salas de aula iluminadas?
E_1_2X	Salas de aula arejadas?
E_1_3X	Ruído prejudica?
E_1_4X	Salas de aula estão limpas?
E_2_1X	Banheiros estão funcionando?
E_2_2X	Banheiros estão limpos?
E_2_3X	Algum banheiro está interditado?
E_6_0X	Avaliação das condições de infra-estrutura da escola
NCOMPX	Número de computadores da escola
BIBLIOT	Existe biblioteca?
LABCIENC	Existe laboratório de ciências?
LABINFO	Existe laboratório de informática?

O arquivo original de escolas da 4ª série do Ensino Fundamental continha 4185 observações, das quais 60 tinham peso amostral igual a zero. Das 4125 observações restantes, 60 observações foram também eliminadas, pois apresentavam valores omissos para todas as perguntas. O banco de

dados utilizado consistiu então de 4065 observações. Esse banco ainda apresentava diversos valores omissos para as diferentes variáveis, mas os mesmos foram considerados como mais uma categoria de resposta e recodificados como o valor mais alto dentro de cada variável. Tal procedimento permitiu que não se perdessem ainda mais observações, já que as técnicas usuais de redução de dimensionalidade não conseguem lidar com valores omissos. Caso eliminássemos também esses, o número de observações cairia para 3857.

Procedimentos semelhantes foram implementados para os dados da 8ª série do primeiro grau e da 3ª série do segundo grau. O tamanho da amostra utilizada foi de, respectivamente, 2558 e 2116 escolas.

As tabelas 2 a 4 apresentam, para cada uma das séries estudadas, os valores possíveis de cada variável, bem como a freqüência absoluta e relativa de cada valor na população, além de outras informações que serão analisadas na seção 4. Saliente-se que o valor mais alto de cada variável corresponde ao valor omisso.

Como se pode notar, as características das escolas não estão distribuídas uniformemente entre as situações possíveis: para alguns quesitos do questionário do SAEB (os relativos ao funcionamento das salas de aula e apoio), com exceção do *e_6_0x* (a medida global e genérica da situação da escola), as instituições de ensino analisadas estão preferencialmente numa situação favorável[1] (ver também Gráfico 1). Além disso, a situação das séries estudadas é muito semelhante com respeito a essas variáveis. Em relação a laboratórios de ciência e informática, estão mais freqüentemente em situação desfavorável. Os quesitos de computadores e biblioteca encontram-se numa situação intermediária. Para tais quesitos retirados do censo escolar, existe um diferencial na amostra com uma melhoria proporcional ao aumento das séries estudadas.

[1] No que diz respeito à variável e_6_0x, considerou-se situação favorável as respostas 4 (bom) e 5 (muito bom). As outras respostas foram consideradas desfavoráveis. No que diz respeito à variável que mede a existência de computadores, considerou-se como desfavorável a não-existência dos mesmos e favorável a existência de qualquer quantidade.

Tabela 2 - Características das escolas e Resultado do GOM
4ª série do Ensino Fundamental

VARIÁVEL	VALOR	CONTAGEM	DISTRI-BUIÇÃO (%)	PERFIL1 (%)	PERFIL2 (%)	DISCRIMI-NAÇÃO
E_1_1X	0 (não)	623	15,30	0,00	28,16	0,158597
	1 (sim)	3440	84,60	100,00	71,84	
	2 (omisso)	2	0,00	0,00	0,00	
E_1_2X	0 (não)	773	19,00	0,00	34,54	0,238602
	1 (sim)	3285	80,80	100,00	65,46	
	2 (omisso)	7	0,20	0,00	0,00	
E_1_3X	0 (não)	2961	72,80	78,77	68,02	0,022836
	1 (sim)	1087	26,70	20,91	31,53	
	2 (omisso)	17	0,40	0,32	0,45	
E_1_4X	0 (não)	421	10,40	0,00	19,08	0,075333
	1 (sim)	3630	89,30	100,00	80,28	
	2 (omisso)	14	0,30	0,00	0,64	
E_2_1X	0 (não)	730	18,00	0,00	33,37	0,222711
	1 (sim)	3304	81,30	100,00	66,63	
	2 (omisso)	31	0,80	0,00	0,00	
E_2_2X	0 (não)	599	14,70	0,00	27,04	0,154854
	1 (sim)	3433	84,50	100,00	71,45	
	2 (omisso)	33	0,80	0,00	1,51	
E_2_3X	0 (não)	2992	73,60	85,32	64,16	0,082968
	1 (sim)	1036	25,50	14,68	34,15	
	2 (omisso)	37	0,90	0,00	1,69	
E_6_0X	1 (péssimo)	108	2,70	0,00	4,96	0,684455
	2 (ruim)	351	8,60	0,00	16,19	
	3 (regular)	1222	30,10	0,00	57,08	
	4 (bom)	1729	42,50	67,68	19,32	
	5 (muito bom)	576	14,20	30,98	0,00	
	6 (omisso)	79	1,90	1,34	2,46	
NCOMPX	1 (não tem)	2849	70,10	44,61	89,82	0,296731
	2 (até 10)	455	11,20	26,07	0,00	
	3 (11 a 100)	224	5,50	12,32	0,00	
	4 (101 e mais)	170	4,20	9,27	0,00	
	5 (omisso)	367	9,00	7,73	10,18	
BIBLIOT	0 (não)	2052	50,50	0,00	100,00	1,951250
	1 (sim)	1966	48,40	97,50	0,00	
	2 (omisso)	47	1,20	2,50	0,00	
LABCIENC	0 (não)	3410	83,90	62,90	100,00	0,257918
	1 (sim)	608	15,00	34,59	0,00	
	2 (omisso)	47	1,20	2,51	0,00	
LABINFO	0 (não)	3191	78,50	48,21	100,00	0,511702
	1 (sim)	827	20,30	49,28	0,00	
	2 (omisso)	47	1,20	2,51	0,00	

Tabela 3 - Características das escolas e Resultado do GOM
8ª série do Ensino Fundamental

VARIÁVEL	VALOR	CONTAGEM	DISTRI-BUIÇÃO (%)	PERFIL1 (%)	PERFIL2 (%)	DISCRIMI-NAÇÃO
E_1_1X	0 (não)	240	9,3	0	14,72	0,051548
	1 (sim)	2308	89,2	100	82,88	
	2 (omisso)	40	1,5	0	2,39	
E_1_2X	0 (não)	398	15,4	0	24,41	0,131922
	1 (sim)	2150	83,1	100	73,21	
	2 (omisso)	40	1,5	0	2,38	
E_1_3X	0 (não)	1802	69,6	77,44	65,06	0,025351
	1 (sim)	741	28,6	22,56	32,19	
	2 (omisso)	45	1,7	0	2,74	
E_1_4X	0 (não)	267	10,3	0	16,36	0,063562
	1 (sim)	2278	88	100	81	
	2 (omisso)	43	1,7	0	2,64	
E_2_1X	0 (não)	339	13,1	0	20,83	0,102375
	1 (sim)	2196	84,9	100	75,93	
	2 (omisso)	53	2	0	3,24	
E_2_2X	0 (não)	370	14,3	0	22,7	0,11917
	1 (sim)	2167	83,7	100	74,18	
	2 (omisso)	51	2	0	3,12	
E_2_3X	0 (não)	1870	72,3	88,19	62,74	0,112241
	1 (sim)	652	25,2	11,81	33,22	
	2 (omisso)	66	2,6	0	4,04	
E_6_0X	1 (péssimo)	40	1,5	0	0	0,416646
	2 (ruim)	141	5,4	0	8,66	
	3 (regular)	626	24,2	0	39,17	
	4 (bom)	1207	46,6	47,7	47,94	
	5 (muito bom)	488	18,9	50,51	0	
	6 (omisso)	86	3,3	1,8	4,22	
NCOMPX	1 (não tem)	1511	58,4	0	89,45	1,113912
	2 (até 10)	348	13,4	39,08	0	
	3 (11 a 100)	278	10,7	30,7	0	
	4 (101 e mais)	231	8,9	25,24	0	
	5 (omisso)	220	8,5	4,98	10,55	
BIBLIOT	0 (não)	574	22,2	0	35,54	0,237949
	1 (sim)	1993	77	97,8	64,46	
	2 (omisso)	21	0,8	2,2	0	
LABCIENC	0 (não)	1755	67,8	20,35	100	1,234901
	1 (sim)	812	31,4	77,46	0	
	2 (omisso)	21	0,8	2,2	0	
LABINFO	0 (não)	1674	64,7	0	100	1,956968
	1 (sim)	893	34,5	97,8	0	
	2 (omisso)	21	0,8	2,2	0	

Tabela 4 - Características das escolas e Resultado do GOM
3ª série do Ensino Médio

VARIÁVEL	VALOR	CONTAGEM	DISTRI-BUIÇÃO (%)	PERFIL1 (%)	PERFIL2 (%)	DISCRIMI-NAÇÃO
E_1_1X	0 (não)	193	9,1	0	17,69	0,074623
	1 (sim)	1891	89,4	100	79,39	
	2 (omisso)	32	1,5	0	2,92	
E_1_2X	0 (não)	287	13,6	0	26,34	0,15819
	1 (sim)	1793	84,7	100	70,38	
	2 (omisso)	36	1,7	0	3,28	
E_1_3X	0 (não)	1491	70,5	80,03	61,34	0,059091
	1 (sim)	586	27,7	19,97	35,1	
	2 (omisso)	39	1,8	0	3,56	
E_1_4X	0 (não)	209	9,9	0	19,15	0,088059
	1 (sim)	1871	88,4	100	77,57	
	2 (omisso)	36	1,7	0	3,28	
E_2_1X	0 (não)	274	12,9	0	25,27	0,151264
	1 (sim)	1798	85	100	70,71	
	2 (omisso)	44	2,1	0	4,02	
E_2_2X	0 (não)	326	15,4	0	30,01	0,207404
	1 (sim)	1746	82,5	100	65,98	
	2 (omisso)	44	2,1	0	4,01	
E_2_3X	0 (não)	1475	69,7	86,76	53,06	0,198771
	1 (sim)	587	27,7	13,24	42,01	
	2 (omisso)	54	2,6	0	4,93	
E_6_0X	1 (péssimo)	34	1,6	0	3,09	0,4324
	2 (ruim)	118	5,6	0	10,61	
	3 (regular)	479	22,6	0	41,85	
	4 (bom)	949	44,8	53,71	37,57	
	5 (muito bom)	460	21,7	46,29	0	
	6 (omisso)	76	3,6	0	6,87	
NCOMPX	1 (não tem)	1007	47,6	0	88,73	1,106478
	2 (até 10)	328	15,5	33,37	0	
	3 (11 a 100)	261	12,3	26,31	0	
	4 (101 e mais)	357	16,9	36,48	0	
	5 (omisso)	163	7,7	3,84	11,27	
BIBLIOT	0 (não)	311	14,7	0	28,75	0,174197
	1 (sim)	1788	84,5	100	69,78	
	2 (omisso)	17	0,8	0	1,47	
LABCIENC	0 (não)	1186	56	17,91	98,44	1,322628
	1 (sim)	913	43,1	82,09	0	
	2 (omisso)	17	0,8	0	1,56	
LABINFO	0 (não)	1116	52,7	0	98,46	1,969674
	1 (sim)	983	46,5	100	0	
	2 (omisso)	17	0,8	0	1,54	

3. Metodologia

Conforme já mencionado anteriormente, utilizou-se o método do "Grade of Membership" (GOM) com a finalidade de reduzir a dimensionalidade das informações sobre as escolas selecionadas na amostra do SAEB-99 no referido estudo sobre proficiência dos alunos e o efeito escola. Essa técnica foi desenvolvida a partir da teoria de conjuntos difusos (Zadeh, 1965) e é utilizada na modelagem multidimensional de dados discretos.

Em contraste com a teoria de conjuntos bem definidos, onde um elemento pertence ou não a um determinado conjunto, na teoria dos conjuntos difusos um elemento pode pertencer parcialmente a vários conjuntos.

Seja um espaço de respostas categóricas com dimensão J. Considerem-se, nesse espaço, elementos i referentes a I observações e K perfis, caracterizadores de situações extremas. Em tal contexto, o grau, segundo o qual um elemento i pertence a um determinado conjunto k, denominado grau de pertinência g_{ik}, é uma variável contínua no intervalo [0; 1]. Os extremos do intervalo têm significados específicos, já que o "0" equivale à situação em que o elemento não pertence ao conjunto, enquanto "1" significa que o elemento pertence única e exclusivamente ao conjunto considerado. Se todos os valores de g_{ik} forem iguais a 0 ou 1, então teremos uma classificação bem definida ao invés de conjuntos difusos.

O grau de pertinência[2] g_{ik} possui as seguintes propriedades:

$$g_{ik} \geq 0 \quad \forall i \ e \ k,$$

$$\sum_{k=1}^{K} g_{ik} = 1 \quad \forall i,$$

Ainda que tome valores no intervalo [0; 1], o grau de pertinência não é uma probabilidade. O grau representa a proporção de intensidade de pertinência ou de proximidade em relação a um determinado conjunto, ou seja, é uma medida matemática de quantidade, que expressa, por exemplo, quantos dos J atributos de um perfil extremo, um determinado elemento possui (Manton, Woodbury, Tolley, 1994). Isto é diferente de uma asserção estatística sobre a probabilidade de pertinência a um dado conjunto, onde os elementos individuais pertencem, *a priori*, a certos conjuntos, e as probabilidades medem o grau de incerteza, dadas as informações disponíveis, sobre tal pertinência.

[2] Que terá, então, o papel da estatística capaz de resumir as informações dos dados multidimensionais, no nosso caso, em uma única escala, já que vamos trabalhar com apenas dois perfis extremos. Esta estatística foi utilizada como covariável no modelo de regressão multinível do estudo já mencionado.

Cabe ressaltar que o grau de pertinência g_{ik} refere-se a uma observação específica *i*. Entretanto, no momento em que se define o número de conjuntos a serem considerados no estudo, faz-se necessário estabelecer uma relação entre os perfis extremos e as categorias das variáveis utilizadas. Isto é feito a partir da estimativa dos parâmetros λ_{kjl}, que representam a probabilidade de uma resposta *l* da *j*-ésima variável para a observação com perfil extremo *k*. Os parâmetros λ_{kjl} prestam-se, assim, para a identificação dos perfis extremos. Um caso particular é quando para uma dada variável λ_{kjl} =0 ou 1 para todos os perfis extremos, *k*, e categorias de variáveis, *l*. Nesse caso, os perfis extremos não tem respostas em comum, são completamente disjuntos e discriminados. No nosso exemplo, a variável *labinfo* apresenta tal comportamento para a 8ª e 11ª séries. No perfil "bom" a escola tem obrigatoriamente laboratório de informática e no perfil "ruim", não. Por outro lado, a variável *e_1_3x* (nível de ruído) não discrimina os perfis em nenhuma das séries, ainda que as escolas no perfil "bom", este seja um problema menos freqüente.

Em suma, a técnica do GOM consiste em estimar, a partir de um modelo de probabilidade multinomial, dois tipos de parâmetros: um de associação de cada elemento *i* ao perfil extremo *k*, g_{ik}; e outro de estrutura, que define as características dos perfis extremos *k*, a partir dos valores *l* tomados pelas variáveis nas *J* dimensões, λ_{kjl}.

Tais parâmetros são obtidos a partir da maximização da verossimilhança do modelo, que pode ser escrito da seguinte forma:

$$L(y) = \prod_{i=1}^{I} \prod_{j=1}^{J} \prod_{l=1}^{L_j} \left(\sum_{k=1}^{K} g_{ik} \lambda_{kjl} \right)^{y_{ijl}},$$

onde *I* é o número de observações na amostra; *J*, o número de variáveis incluídas, ou seja a dimensão do espaço de observações; *Lj*, o número de categorias de cada uma das *J* variáveis; e *K*, o número de perfis extremos.

A maximização é feita iterativamente, otimizando-se com respeito aos parâmetros de estrutura (λ_{kjl}) e de associação (g_{ik}).

4. Resultados

As tabelas 2 a 4 e os gráficos 2 a 9 apresentam, para cada uma das séries, os resultados das estimativas dos parâmetros de estrutura (λ_{kjl}) para cada categoria das variáveis consideradas nos dois perfis obtidos como extremos através do GOM. Além disso, uma medida de discriminação de cada variável, definida como a soma de quadrados das diferenças das

probabilidades em cada perfil, é apresentada na última coluna de cada tabela.

Em todas as tabelas e gráficos, o perfil 1 é o perfil ideal (com o maior número de características boas), e o perfil 2, o não-desejado (com o maior número de características ruins). Note-se que algumas variáveis discriminam com probabilidade 1 um dos perfis, ainda que o conjunto dessas variáveis seja ligeiramente diferente para cada série.

Na 4ª série, para o perfil 1 (bom), as variáveis *e_1_1x*, *e_1_2x*, *e_1_4x*, *e_2_1x*, *e_2_2x*, *e_6_0x* e *bibliot* associam probabilidade 1 a um subconjunto dos valores possíveis, usualmente um único valor, à exceção de *e_6_0x* e *bibliot* (ver também gráficos 2 e 3). Cumpre notar que, para a primeira dessas variáveis, com exceção dos valores omissos, 100% da probabilidade está distribuída em duas categorias, "bom" e "muito bom", hierarquicamente diferenciadas das demais. Para a segunda variável, desconsiderando-se os valores omissos, 100% da probabilidade está concentrada em uma única categoria (ter biblioteca). Para o perfil 2 (ruim), somente as variáveis provenientes do censo escolar (*ncompx*, *bibliot*, *labcienc* e *labinfo*) discriminam perfeitamente. Analisando-se a medida de discriminação que mede a diferença das probabilidades dos dois perfis, vemos que a existência ou não de biblioteca, a avaliação geral das condições de infra-estrutura da escola, bem como a existência ou não de laboratório de informática, são as variáveis que mais discriminam o perfil bom do ruim. A variável que menos discrimina é a *e_1_3x*, que mensura o prejuízo causado por ruído na performance escolar.

Como a distribuição dos valores de cada variável na população estudada apresenta, quase sempre, uma alta concentração num determinado valor (em torno de 80% na maior parte dos casos), os perfis não podem ser disjuntos. Estritamente, ignorando-se os valores omissos, isto só ocorreu para a variável biblioteca (os bons sempre têm biblioteca, enquanto os ruins nunca a possuem). Em menor grau, ainda se observa algo nesta linha na variável *e_6_0x* (avaliação geral das condições de infra-estrutura da escola), onde o perfil bom está sempre avaliado em bom e ótimo, e o perfil ruim, em péssimo, ruim e razoável, além de uma pequena probabilidade em bom (19,32%).

As tabelas 3 e 4 apresentam resultados paralelos para as 8ª e 11ª séries. Conforme pode ser visto, o número de variáveis com categoria associada com probabilidade 1 no perfil bom para ambas as séries (*e_1_1x*, *e_1_2x*, *e_1_4x*, *e_2_1*x, *e_2_2x*, *e_6_0x*, *bibliot*, *labinfo*) é ligeiramente mais expressivo nas referidas séries do que na 4ª série (uma variável a mais, *labinfo*). Por outro lado, o número de variáveis com categoria associada com probabilidade 1 no perfil ruim para ambas as séries diminui (*labcienc* não associa mais probabilidade a nenhuma categoria específica). Nessas duas

séries, existem duas variáveis, *ncompx* e *labinfo*, que discriminam perfeitamente os perfis bons dos ruins. Outra variável que apresenta um alto grau de discriminação nas citadas séries é a *labcienc*. A que menos discrimina é a mesma observada na 4ª série, *e_1_3x*. A variável *e_6_0x*, que classifica de forma global as condições de infra-estrutura da escola, já não é mais tão importante para discriminar, possivelmente pelo fato de que a distribuição entre as diferentes classes é mais uniforme nas séries mais avançadas do que na 4ª série.

Gráfico 1

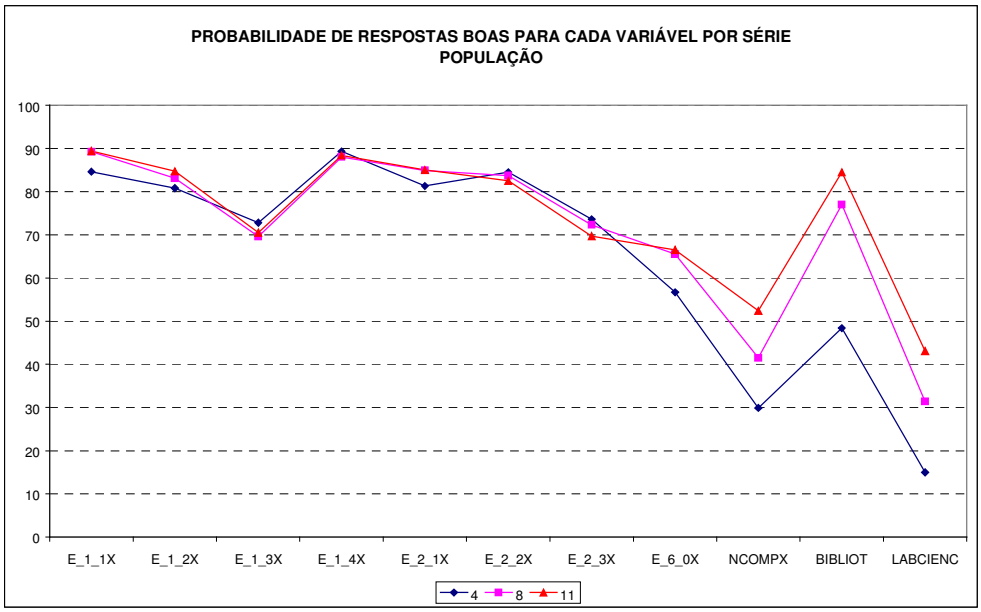


Gráfico 2

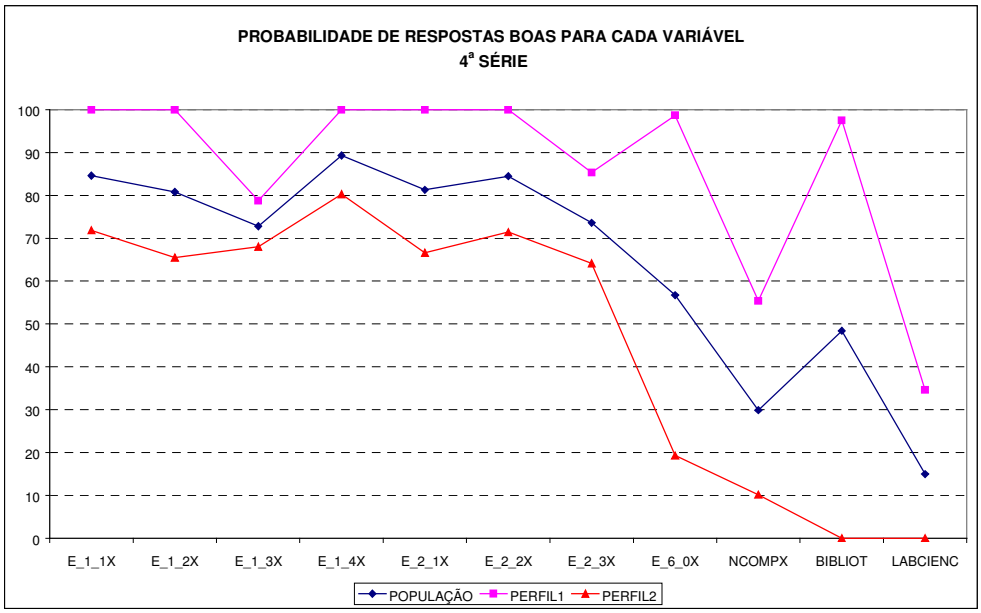


Gráfico 3

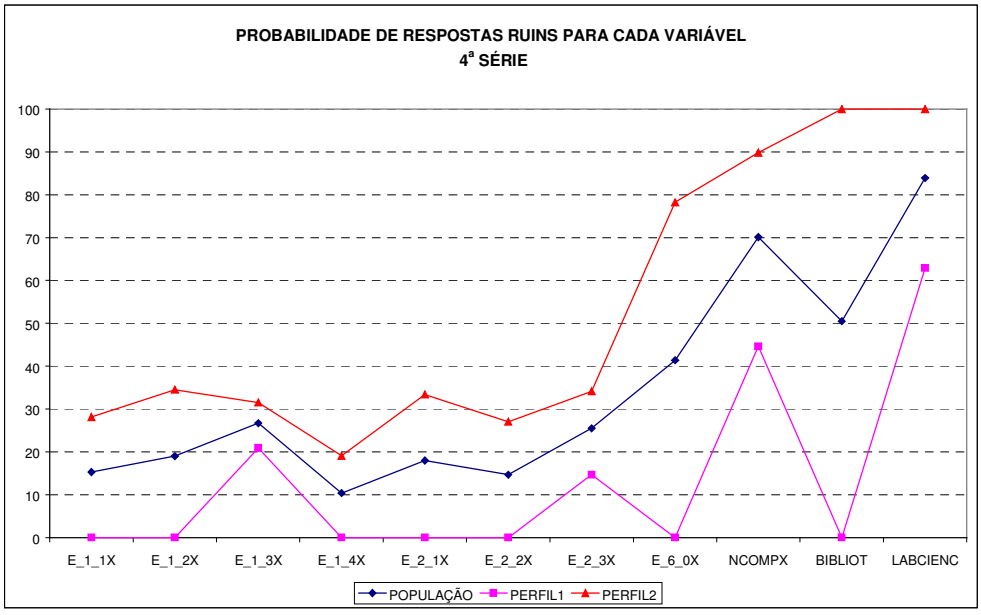


Gráfico 4

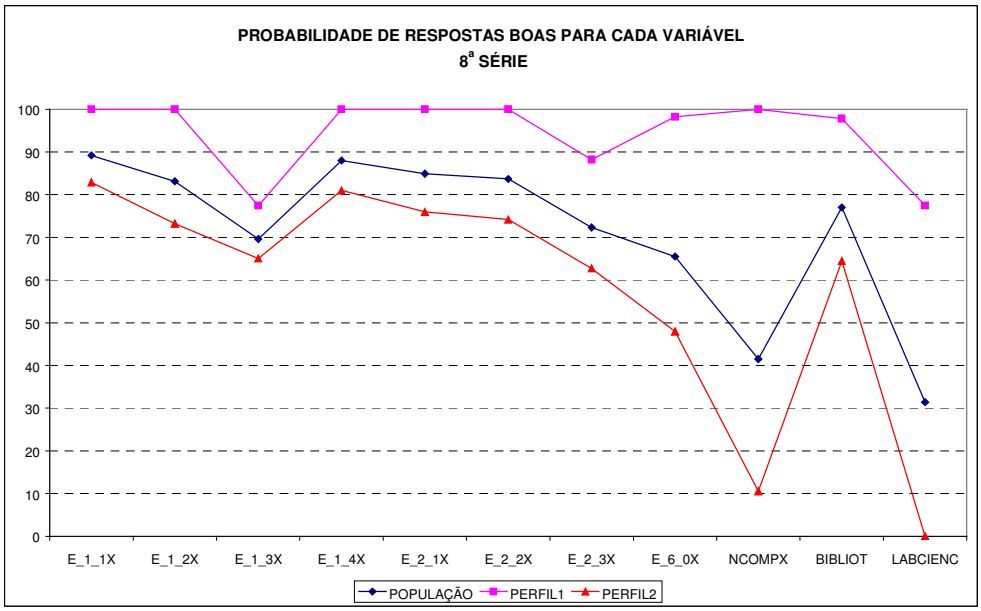


Gráfico 5

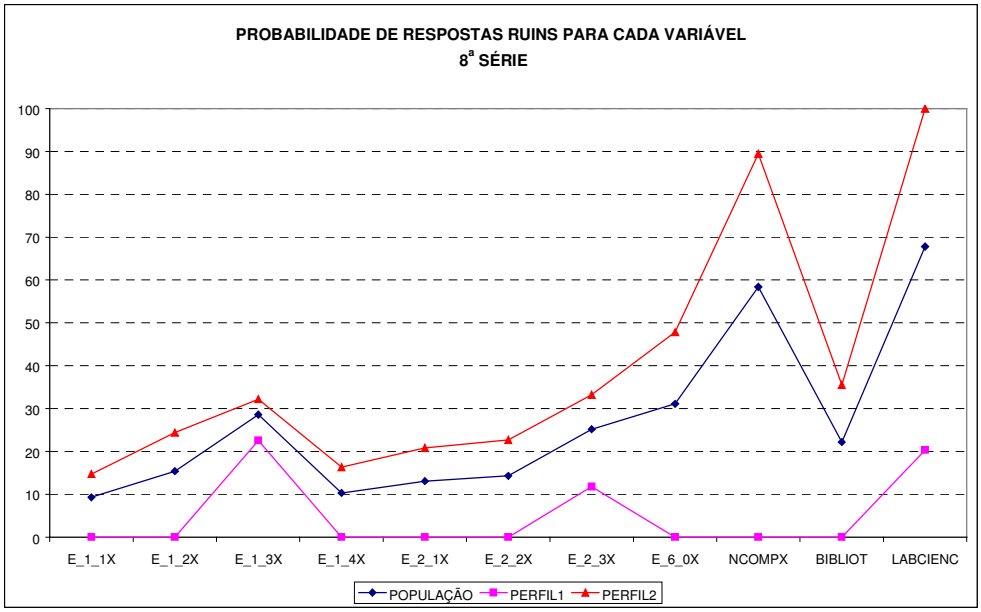


Gráfico 6

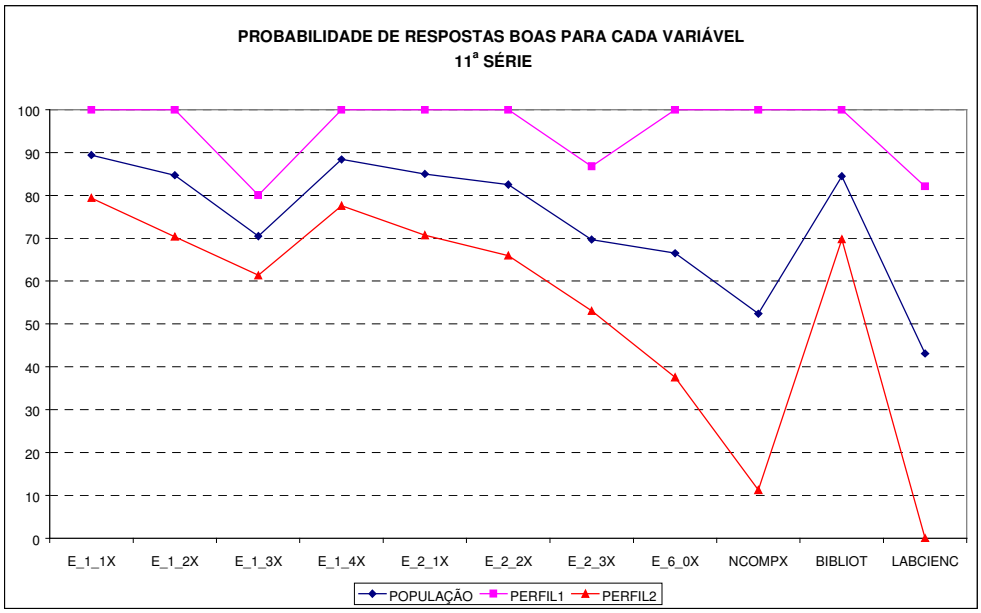


Gráfico 7

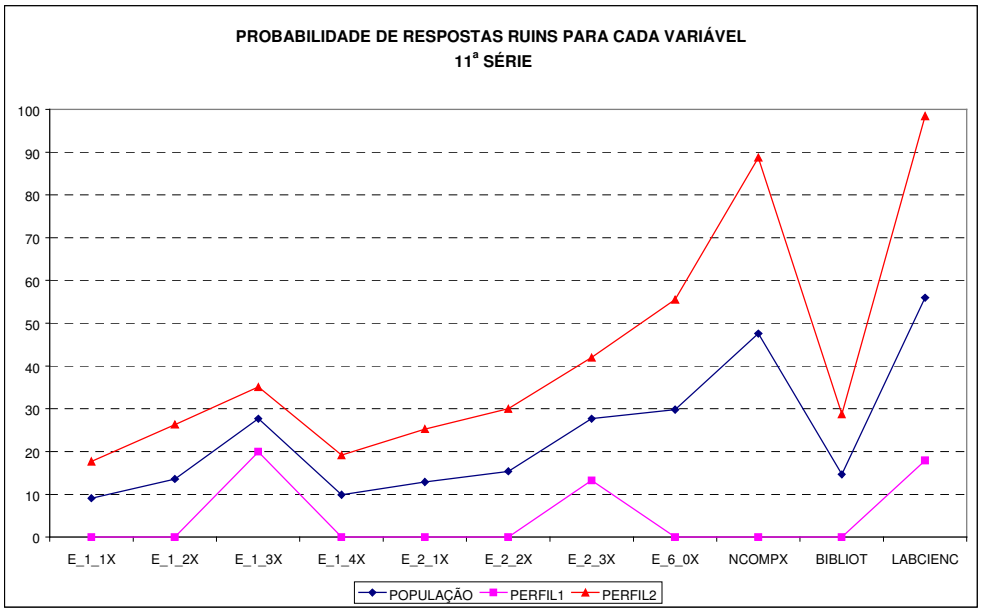


Gráfico 8

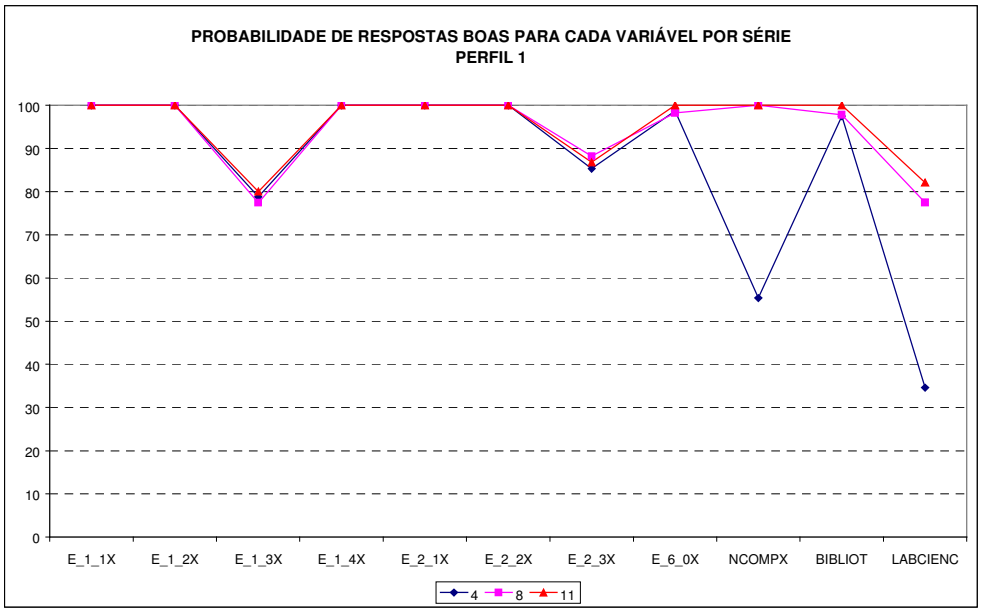
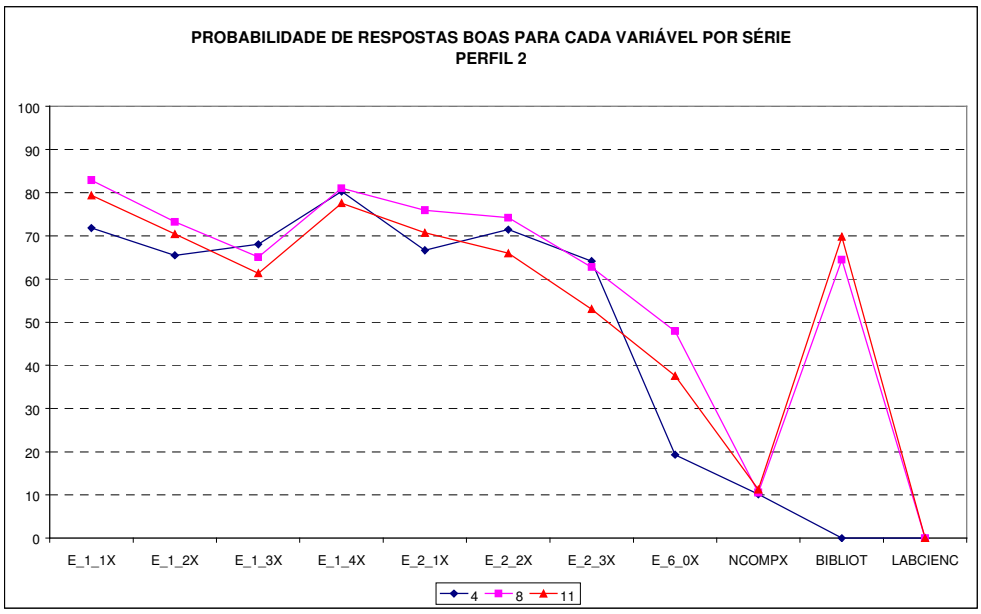


Gráfico 9



5. Conclusão

Neste estudo, a técnica do GOM mostrou-se apropriada para a redução da dimensionalidade dos dados referentes às características das escolas. Dois perfis foram gerados: um para a escola boa e um para a escola ruim. O perfil da escola boa é bem específico. O da escola ruim é mais difuso. Ao final do processo de análise, graus de pertinência, que associam cada elemento ao melhor perfil foram gerados e incorporados ao banco de dados original, os quais foram utilizados na aplicação dos modelos de regressão multinível. A variável criada a partir dessa técnica mostrou-se um forte preditor da proficiência nos modelos utilizados.

As variáveis da escola referentes à segurança e violência não foram analisadas. Um possível desdobramento deste estudo seria a incorporação dessas variáveis para enriquecer os perfis utilizados. Poder-se-ia também testar a expansão da classe de perfis extremos.

6. Referências Bibliográficas

BARBOSA, M. E. F.; FERNANDES, C.; SANTOS, D.; LEITE, I. C.; BELTRÃO, K. I.; FARIÑAS, M. *Análise descritiva dos dados do SAEB-99*, Relatório técnico INEP/MEC, mimeo, 2000a.

BARBOSA, M. E. F.; FERNANDES, C.; SANTOS, D.; LEITE, I. C.; BELTRÃO, K. I.; FARIÑAS, M. *Redução da dimensionalidade dos dados do SAEB-99*, Relatório técnico INEP/MEC, mimeo, 2000b.

BARBOSA, M. E. F.; FERNANDES, C.; SANTOS, D.; BELTRÃO, K. I.; FARIÑAS, M. *Análise exploratória dos dados provenientes do censo escolar 1999, acrescentados ao SAEB*, Relatório técnico INEP/MEC, mimeo, 2000.

BARBOSA, M. E. F.; FERNANDES, C.; SANTOS, D.; BELTRÃO, K. I.; FARIÑAS, M. *Modelos Multinível*, Relatório técnico INEP/MEC, mimeo, 2001.

MANTON, K. G.; WOODBURY, M. A.; TOLLEY, H. D. *Statistical Applications Using Fuzzy Sets,* John Wiley & Sons, Inc., New York, 1994.

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

ZADEH, L. A. Fuzzy sets. *Information Control* , n. 8, p. 338-353, 1965.