6
Alternativas aos Modelos Clássico da TRI

Após todas as análises, fica evidente que se necessita de um especialista para se estimar um indicador de nível sócio econômico usando a teoria de reposta ao item. Na TRI existem estatísticas muito complexas e os softwares existentes hoje para o usuário não possuem um ambiente muito amigável no seu manuseio, como é o caso dos softwares: BILOG-MG, PARSCALE, etc. Por outro lado, na interpretação da escala obtida o número de classes ou agrupamentos de classificação dos indivíduos deve ser definido subjetivamente pelos especialistas, pois esse tipo de modelo não fornece um numero ótimo de classes estimadas (pelo menos *a priori*), embora seja possível se utilizar de técnicas clássicas de *cluster analysis* como auxílio à classificação. No presente estudo trabalhamos com dez classes para classificação. Porém nada nos impede de trabalharmos com um número maior ou menor que esse. Na verdade até podemos buscar um número ótimo, através do emprego de análises exaustivas. Note-se que, quanto maior o número de classes criadas na interpretação da escala, mais incerta fica a classificação de um indivíduo, pois devido ao erro de estimação pode estar em uma classe ou em outra.

É bem verdade que o emprego da TRI na produção de um indicador sócio-econômico será um motivador para a construção de muitos outros tipos de *constructos latentes* em áreas afins, devido a grande potencialidade da TRI. Já a comparação com o critério adotado atualmente pelos pesquisadores de marketing e outros foi conseqüência natural do trabalho, pois entendemos que um trabalho isolado sem comparabilidade fica sem sentido. Embora tenhamos ressaltado as limitações do critério atualmente empregado e as grandes vantagens que a TRI proporciona, devemos tomar cuidado com as conclusões. Apenas mostramos que é bastante viável a construção desse tipo de indicador a partir do emprego de modelos de traços latentes. Entendemos que a opção em adotar este ou àquele critério deve levar em conta além dos critérios técnicos a familiaridade com o tipo de ferramenta e a simplicidade de seu emprego.

Será que realmente é necessária uma teoria tão formal para a classificação de indivíduos em estratos sócio-econômicos? Essa dúvida é pertinente, sobretudo, no contexto de que há necessidade de uma metodologia que conduza à produção do indicador de forma mais simples e rápida. Nesse sentido, perguntamo-nos se existe uma teoria com vantagens similares ao da TRI, porém com *softwares* muito mais amigáveis para o seu manuseio e que permitam métodos para a produção do indicador mais fáceis de serem utilizados. Acreditamos que talvez essa resposta possa ser alcançada com o emprego de modelos de classes latentes.

A TRI tradicional emprega modelos de traços latentes porque o *constructo* latente é considerado como uma variável numérica contínua. Num modelo de classes latentes o *construto* é considerado como uma variável discreta, categorias nominais ou ordinais, segundo as quais os indivíduos são classificados. Pode-se ter, ainda, um número *ótimo* de grupos (ou categorias) estimados ao se adotar um critério estatístico para comparação entre os diversos modelos naturalmente aninhados (como por exemplo, as estatísticas AIC e BIC). Além disso pode-se estimar as respectivas probabilidades de um certo indivíduo pertencer aos diferentes grupos considerados.

Como já ressaltado anteriormente, uma desvantagem da TRI tradicional é a dificuldade para não *experts* em produzir uma estimativa do traço latente devido a complexidade (para o não especialista) em conduzir a estimação de verossimilhança máxima (ou, mesmo dos métodos Bayesianos). Essa dificuldade pode ser contornada no caso dos modelos de classes latentes, bastando fornecer as regras de classificação a partir das respostas atribuídas às variáveis indicadoras, o que já é fornecido automaticamente em determinados *softwares*. Além disso, *softwares* comerciais permitem construir modelos que admitem diferentes escalas de medidas para as variáveis indicadoras em um só questionário. Daremos a seguir uma pequena introdução na teoria de modelos de classes latentes e depois mostraremos alguns resultados experimentais com as mesmas bases da amostra usada no trabalho, e apresentaremos uma comparação entre os resultados dos três modelos citados.

6.1
Modelos de Classe Latente para análise de agrupamentos (L C Cluster)

Modelos de classe latente para análises de agrupamento foram propostos por Lazarsfeld & Henry (1968), Goodman (1974) e Haberman (1979), originalmente admitiam variáveis indicadoras dicotômicas, mas atualmente permitem misturar e aplicar modelos que incluem variáveis com tipos de escalas diferentes (nominal, ordinal, contínua e de contagem) além de incluírem covariáveis (Clogg, 1981, Hagenaars, 1990).

Nesse tipo de modelo admite-se que a variável latente θ seja uma variável ordinal ou nominal (daí decorre o termo classe latente), diferente dos modelos comuns da TRI onde a variável latente é numérica e contínua (modelos de traço latente). Especificamente, no caso de modelos para análise de grupamento, uma única variável latente é considerada (caso unidimensional, portanto). Admite-se, ainda, que existe um conjunto de variáveis indicadoras $y = (y_1, y_2, \ldots, y_I)$, que podem ser nominais , ordinais, numéricas contínuas ou numéricas de contagem. Finalmente, admite-se um conjunto de covariáveis independentes $z = (z_1, \ldots, z_L)$, que podem ser nominais ou numéricas (ordinais, continua ou discretas).

Se y_i for nominal ou ordinal, o modelo que associa y_i à variável latente θ e às covariáveis é dado por:

$$f_i(y;\theta,z) = P(y_i = k;\ \theta, z) = \frac{e^{(\eta^i_{k;\theta,z})}}{\sum_{k=1}^{m_i} e^{(\eta^i_{k;\theta,z})}}$$

onde $\eta^i_{k;\theta,z} = \beta^0_{k_i} + \beta^1_{k_i,\theta} + \sum_{j=1}^{L} \beta^2_{k_i,Z_j}$, $\theta = \theta_1, ..., \theta_c$, $k_i = 1, ..., m_i$.

Tipicamente, $\beta^2_{k_i,z_j} = 0$, hipótese que implica em não existir efeito direto das covariáveis sobre as variáveis indicadoras. Adimite-se, ainda, que se y_i for

ordinal $\beta^1_{k_i,\theta} = \beta^1_{i,\theta}\ v_{k_i}$, onde v_{k_i} é um *score* fixo, corresponde a cada categoria do item i, tal que $v_{k_i} - v_{k_i-1}$ é constante. São restrições necessárias para a identificabilidade do modelo que $\sum_{k_i=1}^{m_i} \beta^0_{k_i} = 0$, $\sum_{k_i=1}^{m_i} \beta_{k_i,\theta} = 0, \forall\, \theta = 1,...,c$ e $\sum_{\theta=1}^{c} \beta_{k_i,\theta} = 0\ \ \forall k_i$ caso y_i seja nominal e $\sum_{\theta=1}^{c} \beta^1_{i,\theta} = 0$ caso y_i seja ordinal.

Se y_i for contínua admite-se que, o modelo proposto será definido a partir de uma distribuição normal:

$$f_i(y;\theta,z) = \frac{1}{\sqrt{2\pi}\ \sigma^i_\theta} \exp\left(\frac{-\frac{1}{2}\left(y - \eta^i_{\theta,Z}\right)^2}{\left(\sigma^i_\theta\right)^2} \right)$$

e, $\eta^i_{\theta,z} = \beta^0_i + \beta^1_{i,\theta}$ desconsiderando-se o efeito direto das covariáveis sobre y_i.

O efeito das covariáveis sobre θ é modelado da seguinte forma:

$$P(\theta = c;\ z) = \frac{\exp(\eta_{c,z})}{\sum_{c=1}^{c} \exp(\eta_{c,z})}$$

onde $\eta_{c,z} = \gamma^0_c + \sum_{l=1}^{L} \gamma^1_{c,z_l}$.

Se θ for ordinal então: $\gamma^0_c = \gamma^0\, v_c$, onde v_c representa a categoria associada a $\theta = c$, tal que $v_c - v_{c-1}$ é constante. Se a covariável for numérica (contínua ou discreta) então $\gamma^1_{c,z_l} = \gamma^1_{c,l}\, z_l$.

Sob os modelos propostos acima, tem-se que a representação marginal (sobre θ) de y_i é dada por:

$$f(y_{I(q)};z) = \sum_{c=1}^{c} P(\theta = c;\ z) \prod_{i \in I(j)} f_i(y;\theta,z),\ ,$$

onde $I(q) \subset (1,...,I)$, admitindo-se independência condicionada a θ dos itens.

Supondo que existam Q padrões diferentes de respostas cada qual com $\overline{r}_q$ casos tem-se que o logaritmo da função de verossimilhança é dado por:

$$L(\mathrm{P}\,; y, z) = \sum_{q=1}^{Q} \overline{r}_q \, \log f(y_{I(q)}; z, \mathrm{P})$$

onde P representa o conjunto de parâmetros do modelo a ser estimado. O software *Latent Gold* para evitar falta de condicionamento numérico e não existência de solução dos estimadores de máxima verossimilhança admite, ainda, a utilização de um método Bayesiano. A idéia básica consiste em admitir distribuições a priori $h(\mathrm{P})$ para os parâmetros e maximizar o logaritmo da distribuição a posteriori:.

$$P_M(\mathrm{P}\,; y, z) = \sum_{q=1}^{Q} \overline{r}_q \, \log f(y_{I(q)}; z, \mathrm{P}) + \log h(\mathrm{P})$$

Detalhes sobre as distribuições *a priori* utilizadas podem ser encontrados em (Magidson, p. 164). As priors são construídas de tal forma que uma constante (*Bayes Constant*) introduzida, a ser especificada pelo usuário, determina o grau de influência da distribuição a priori sobre o resultado da estimação dos parâmetros. O método utilizado no *Latent Gold*® para otimizar o funcional acima é o mesmo baseado no emprego do algoritmo EM de forma bastante similar ao método utilizado na TRI para traços latentes.

6.2
Resultados no uso do LC cluster

Podemos adotar como critério de seleção das variáveis significativas para a construção do modelo de classes latentes o valor R^2, que é um valor similar ao da análise de regressão. Foram escolhidas aquelas variáveis com R^2 maior que 10 %.

Uma maneira de escolher qual o número mais adequado de grupos a formar pode ser pela menor estatística BIC e AIC.

Iniciou-se com a formação de três grupos. Porém, seus resultados não foram satisfatórios. A tentativa e erro neste momento é a única saída na busca da estabilidade do modelo final. A menor estatística BIC e AIC pode nortear o especialista quando o número ótimo de classes estimadas foi encontrado. Aqui adotamos também este critério. E o número de grupos com mais estabilidade e apresentação de resultados mais satisfatórios foram sete grupos. Segue abaixo as variáveis significantes para o modelo. As variáveis que tiveram um R^2 abaixo do valor estipulado, 10 %, foram retiradas de qualquer análise futura.

ITEM	R²	ITEM	R²
Banheiro	0,4099	Maq. Lavar Roupa	0,2548
Carro	0,4584	Piso de cimento	0,2365
Tv	0,3388	Cobert. de cimento	0,5646
Ventilador	0,2019	Cobert. de telha de barro	0,6315
Instrução	0,3166	Cobert. de telha de amianto	0,6154
Área	0,3422	Microondas	0,3279
Paredes	0,1453	Liquidificador	0,289
Empregada	0,3354	Batedeira	0,2438
Freezer	0,2273	Aspirador de pó	0,165
Som	0,1633	Aquecedor	0,7823
Vídeo	0,3438	Geladeira	0,2271
Computador	0,2746	Ar condicionado	0,283
Ferro	0,2282	Chuveiro elétrico	0,5191

Tabela 17 - Resultado dos valores de significância das variáveis para o modelo LC cluster

No anexo 3 apresenta-se a tabela com as probabilidades de um indivíduo com determinada posse de bens pertencer a cada um dos clusters.

A interpretabilidade da tabela no anexo 3 é justamente dizer qual é o perfil (suas posses e grau de instrução) de um indivíduo que tem maior chance de estar em determinado grupo. Assim podemos notar, por exemplo, que o grupo seis é caracterizado por pessoas quem têm maior chance de não ter praticamente

nenhum item dos listados acima; ser semi-analfabeto e a estrutura da casa ser precária. Para ajudar na leitura de todos os grupos formados bem como o que caracteriza tais grupos (com base na tabela 17 do anexo 3), listamos abaixo a sua leitura.

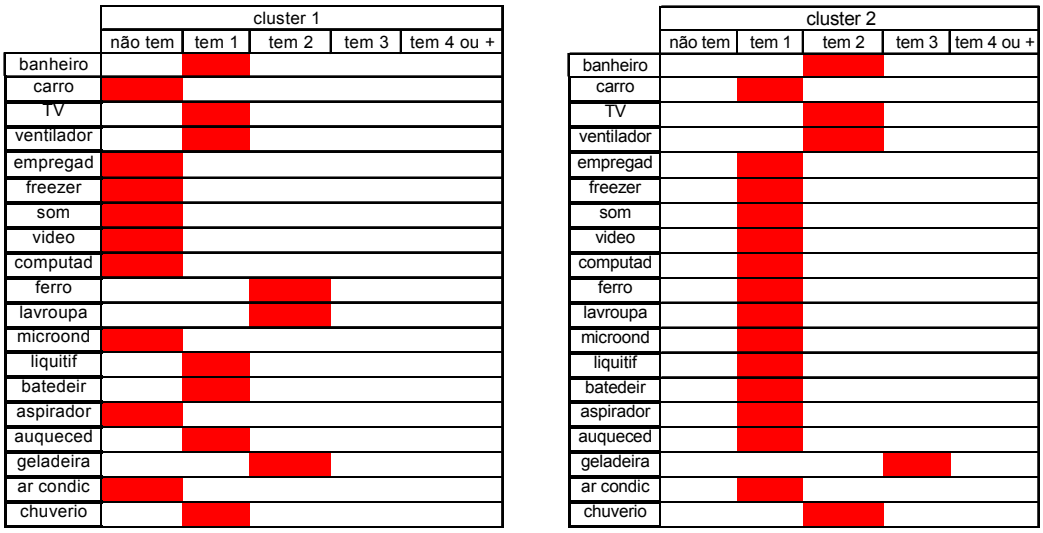
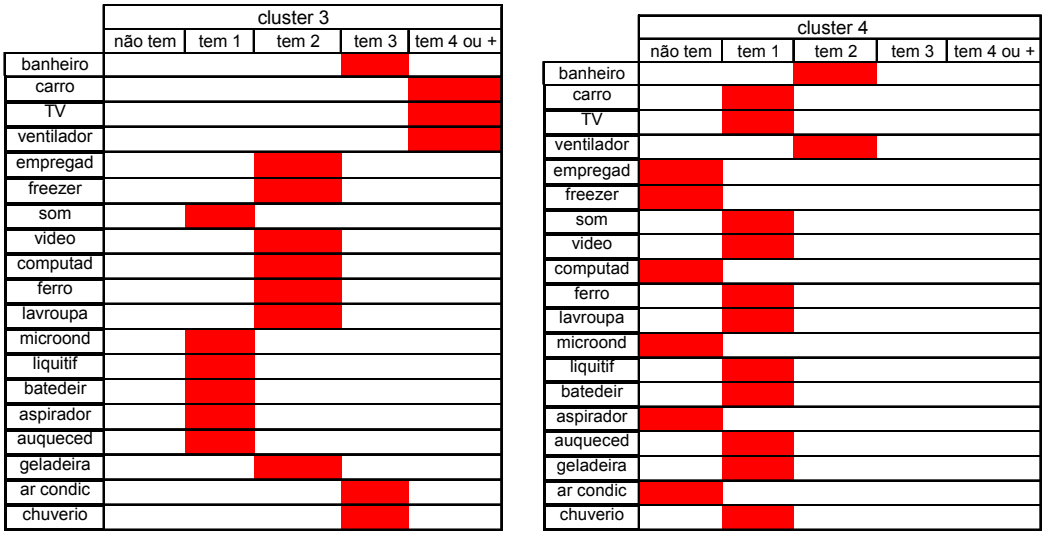


Figura 14 - Característica das posses de um indivíduo que pertence a um determinado cluster



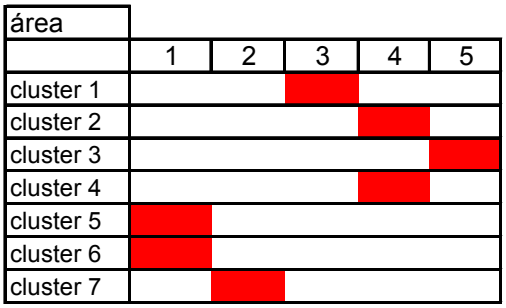
	cluster 5				
	não tem	tem 1	tem 2	tem 3	tem 4 ou +
banheiro		■			
carro	■				
TV		■			
ventilador		■			
empregad	■				
freezer	■				
som	■				
video	■				
computad	■				
ferro		■			
lavroupa	■				
microond	■				
liquitif		■			
batedeir	■				
aspirador	■				
auqueced		■			
geladeira		■			
ar condic	■				
chuverio		■			

	cluster 6				
	não tem	tem 1	tem 2	tem 3	tem 4 ou +
banheiro	■				
carro	■				
TV	■				
ventilador	■				
empregad	■				
freezer	■				
som	■				
video	■				
computad	■				
ferro	■				
lavroupa	■				
microond	■				
liquitif	■				
batedeir	■				
aspirador	■				
auqueced	■				
geladeira	■				
ar condic	■				
chuverio	■				

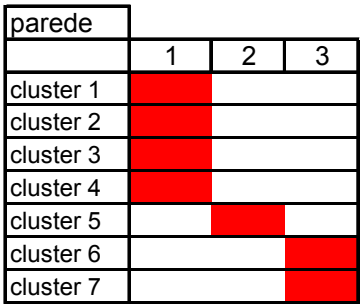
	cluster 7				
	não tem	tem 1	tem 2	tem 3	tem 4 ou +
banheiro	■				
carro	■				
TV	■				
ventilador		■			
empregad	■				
freezer	■				
som	■				
video	■				
computad	■				
ferro		■			
lavroupa	■				
microond	■				
liquitif		■			
batedeir	■				
aspirador	■				
auqueced	■				
geladeira		■			
ar condic	■				
chuverio	■				

instrução					
	1	2	3	4	5
cluster 1	■				
cluster 2				■	
cluster 3				■	
cluster 4					■
cluster 5	■				
cluster 6	■				
cluster 7	■				

1 - até primário incompleto
2 - até ginasial incompleto
3 - até colegial incompleto
4 - até superior incompleto
5 - curso suprior completo



1- até 50 m^2
2 -de 51 à 75 m^2
3 - de 76 à 100 m^2
4 - de 101 à 150 m^2
5 - acima de 151 m^2



1 -alvenaria
2 - madeira
3 - material aproveitado

O modelo de classes latentes ainda permite construir uma tabela com todas as possíveis possibilidades do respondente pertencer aos grupos formados de acordo com suas posses de itens. A tabela abaixo exemplifica com uma amostra de nove respondentes tal afirmação:

	Pessoa 1	Pessoa 2	Pessoa 3	Pessoa 4	Pessoa 5	Pessoa 6	Pessoa 7	Pessoa 8
Banheiro	0	1	1	2	2	2	2	.
Carro	0	0	0	0	1	1	1	.
Tv	0	0	1	1	1	2	2	.
ventilador	0	0	2	1	1	0	2	.
Instrução	1	1	3	3	1	5	3	.
Área	2	1	3	3	3	3	4	.
Parede	2	1	1	1	1	1	1	.
empregada	0	0	0	1	0	1	0	.
Freezer	0	0	0	0	0	1	0	.
Som	0	0	1	1	1	1	1	.
Vídeo	0	0	1	0	0	1	1	.
computador	0	0	0	0	0	0	0	.
Ferro	0	1	1	1	1	1	1	.
Lav.roupa	0	0	1	0	1	1	1	.
Pisoc	1	0	1	0	1	1	1	1
Coblc	0	1	1	0	0	0	0	.
Cobtb	0	0	0	0	0	1	0	.
Cobta	0	1	1	1	1	1	1	.

microondas	0	0	0	0	0	0	0	.
Liquidific	0	1	1	1	1	1	1	.
Batedeira	0	0	1	0	1	0	0	.
aspirador	0	0	0	0	0	0	0	.
aquecedor	0	0	0	1	1	1	0	.
Geladeira	0	0	1	1	1	1	1	1
Ar cond.	0	0	0	0	0	1	0	.
Chuveiro	0	0	0	1	1	1	0	.
Modal	**6**	**6**	**7**	**4**	**4**	**3**	**4**	**3**
Cluster1	0	0	0	0,295	0,078	0	0	0,1906
Cluster2	0	0	0	0,323	0,3759	0,4919	0,13	0,2011
Cluster3	0	0	0	0	0,0007	0,5081	0	0,2025
Cluster4	0	0	0,4409	0,382	0,5454	0	0,46	0,1921
Cluster5	0	0	0	0	0	0	0	0,0957
Cluster6	1	0,9	0	0	0	0	0	0,0235
Cluster7	0	0,1	0,5591	0	0	0	0,4	0,0945

Tabela 18 - Possibilidades de pertinência a um determinado cluster

A partir da tabela descrita acima podemos perceber que de acordo com as posses do respondente um, ele tem 100 % de chance de estar no grupo seis. O respondente três possui aproximadamente 56 % de chance estar no grupo sete. O exemplo do respondente nove é de interesse particular, haja visto que seu caso possui "*missing cases*", assim o mesmo foi classificado no grupo três.

Abaixo podemos visualizar os sete grupos formados pelo modelo de classes latentes. Percebe-se claramente a distinção entre os sete grupos formados. Esta comparabilidade foi feita com o *score* estimado via TRI. Entende-se que os respondentes que estão no cluster três é similar a dizer que os mesmos possuem *score* entre um e dois na TRI. Percebe-se também que o grupo das pessoas com menor potencial de consumo são aqueles que se situam no grupo seis com *score* (TRI) próximo de -2.

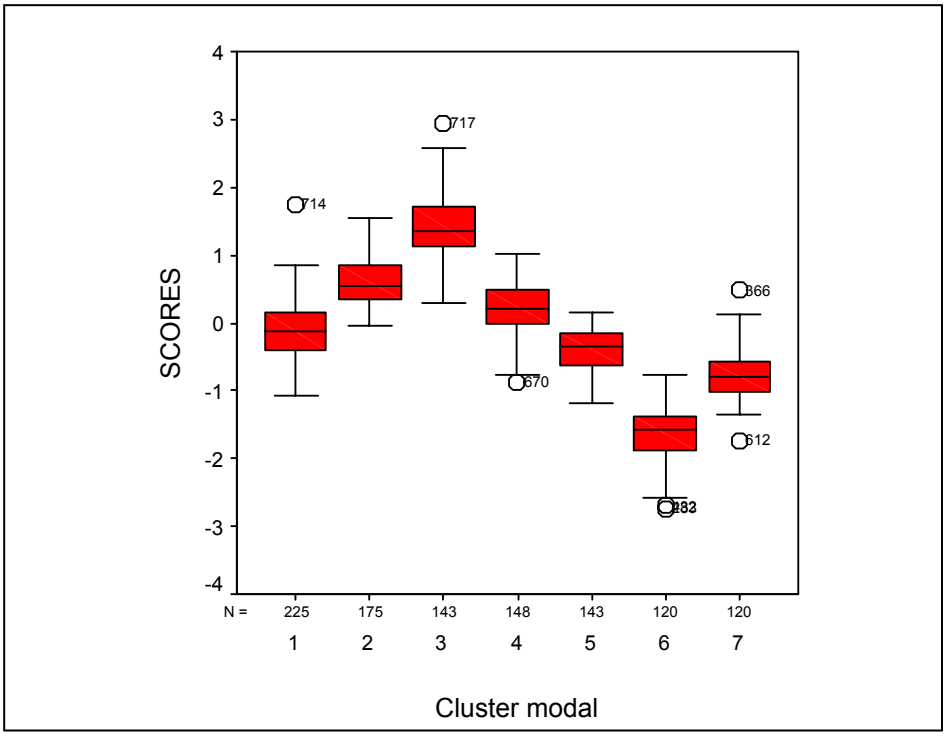


Figura 15 - LC cluster *versus* TRI

Abaixo podemos visualizar os sete grupos formados pelo modelo de classes latentes. Percebe-se claramente a distinção entre os sete grupos formados. Esta comparabilidade foi feita com a pontuação do Critério Brasil. Entende-se que os respondentes que estão no cluster três são aqueles que no Critério Brasil possuem maior pontuação quanto a posse de seus itens. Percebe-se também que o grupo das pessoas com menor potencial de consumo são aqueles que se situam no grupo seis, comparados ao Critério Brasil são as pessoas que pertencem a classe E e D.

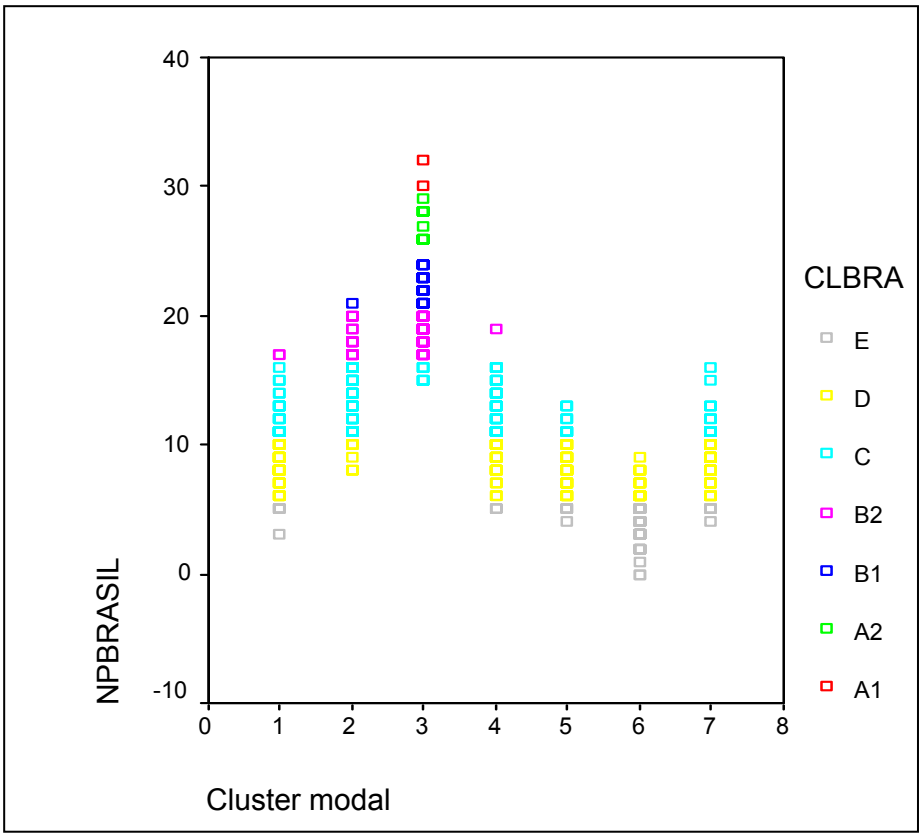


Figura 16 - LC cluster *versus* Critério Brasil

A título de comparabilidade entre os três modos diferentes de classificar um indivíduo em estratos sócios econômicos deixamos registrado as três classificações: Critério Brasil; *score* via TRI e L C cluster.

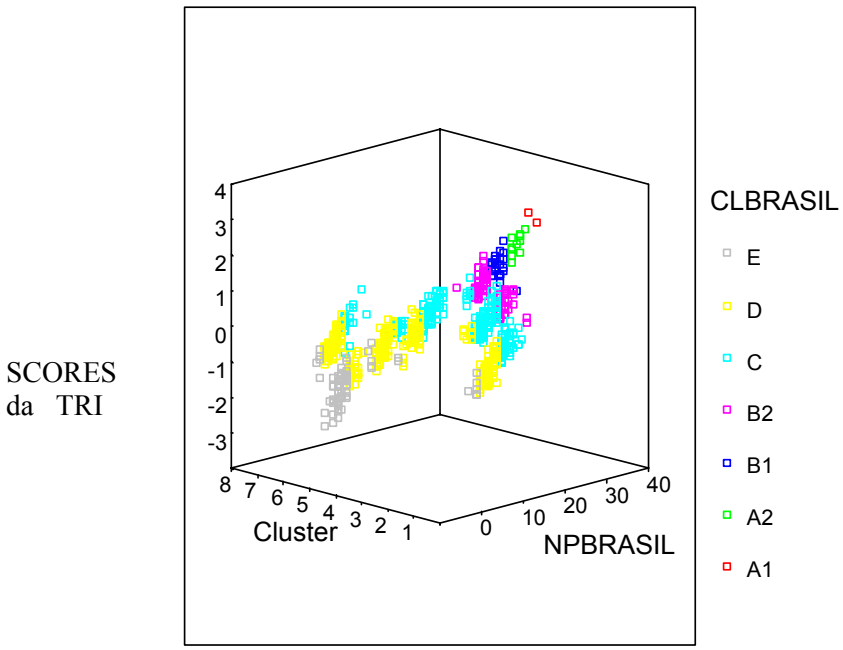


Figura 17 - Os três critérios

CLBRASIL é o critério de classificação Brasil
SCORE via TRI Sá os valores estimados para cada respondente utilizando a TRI.
Clusters são os grupos formados pelo modelo de classes latentes
NPBRASIL é a pontuação do Critério Brasil adotada pelo mesmo.

A figura 17 acima possibilita uma visualização tridimensional dos três critérios citados no presente trabalho. Podemos classificar tal figura como um resumo das figuras anteriores apresentadas no desenvolvimento da dissertação. Embora, a figura 17 não forneça novas informações ao trabalho, achamos interessante mostrar ao leitor a comparação dos três critérios em um só gráfico.