ANÁLISE FUZZY CLUSTER EM AMBIENTE R: UMA APLICAÇÃO DOS ALGORITMOS FANNY, CMEANSnE C-MEDOIDS NA CLASSIFICAÇÃO DOS MUNICÍPIOS BRASILEIROS SEGUNDOINDICADORES DE BEM-ESTAR-SOCIAL

Rodrigo Nunes Ferreira[♣]

Resumo

O artigo demonstra o uso da análise fuzzy cluster através do uso de três algoritmos fanny, cmean e c-medoids na classificação dos municípios brasileiros segundo indicadores de bem estar- social. São utilizados índices de validação de agrupamentos para comparar os três algoritmos utilizados quanto a capacidade de gerar agrupamentos fuzzy com coesão interna e isolamento dos grupos formados. Como estratégia complementar de validação foi gerada uma análise discriminante que busca reproduzir o modelo de agrupamento gerado pela análise de cluster e assim medir o grau de confiabilidade, ou qualidade, do esquema classificatório. Todos os cálculos foram operacionalizados no ambiente R, fazendo uso dos recursos disponíveis na biblioteca do ambiente. Como indicador de bem estar social dos municípios brasileiros foram selecionados os sete indicadores que compõe o IDH-M acrescidos de três indicadores: percentual de pessoas acima da linha de pobreza, percentual de domicílios em condições adequadas de moradia e percentual de pessoas que demoram até trinta minutos no deslocamento casatrabalho. Os resultados mostraram melhores resultados para o algoritmo c-means, que não estáimplementado nos pacotes estatísticos mais utilizados em estudos socioeconômicos. O que mostra a importância do uso dos recursos do ambiente R em estudos desta natureza.

Palavras-chave:análise de agrupamento, fuzzy cluster, validação de agrupamentos, fanny, c-means, cmedoids

[♣]Doutorando do Programa de Pós-Graduação em Geografia/IGC/UFMG.

1-INTRODUÇÃO

Em 1930 o matemático polonês Jan Lukasiewicz desenvolveu os princípios modernos da lógicanebulosa. Na década de 1950 A. Kaplan e H. F. Schott formalizam os princípios da álgebra dos conjuntos nebulosos (fuzzy), partindo do princípio de que os processos de decisão são marcados, muitas vezes, pela "incerteza" presente nos fatores a serem analisados, podendo afetar a exatidão das respostas e a coerência das ações a eles relacionados. A lógica fuzzy, ao utilizar a linguagem matemática - grau de pertinência - associada à representações linguísticas de agrupamentos, fundamenta modelagens como instrumentos de resolução de processos decisórios com fatores de incerteza. Nos conjuntos fuzzy as fronteiras não são nitidamente definidas, um elemento xi poderá pertencer, com certo grau de pertinência, a conjuntos diferentes. O grau de pertencimento é dado por uma função "grupo-conceito", que imputa, em relação a esse "grupo-conceito", graus de pertencimento no intervalo [0,1] a cada elemento xi do universo de discurso. As funções de pertencimento (u) podem ter forte conteúdo subjetivo em sua definição. Assim, um determinado "grupo-conceito" poderá ter diferentes funções de pertencimento; ou, uma função padrão (Jané,2004; Oliveira, 2003; Cruz, 2004)

A teoria dos conjuntos fuzzy foi proposta por Zadeh (1965), que desenvolveu o conceito deincerteza de pertença descrito por uma função de pertinência. As primeiras aplicações da teoria dos conjuntos fuzzy em análise de cluster surgiram nos trabalhos de Bellman, Kalaba e Zadeh (1966) e Ruspini (1969), que desenvolveram o algoritmo FANNY. A concepção de Dunn (1974), aprimorada por Bezdek (1981), permitiu o desenvolvimento do algorítimo Fuzzy C-Means, como uma versão fuzzy do algoritmo K-means de particionamento rígido apresentado por J. MacQueen (1967) (Yang, 1993). Atualmente existe uma grande número de algoritmos fuzzy cluster disponíveis na literatura, em sua maioria variações e aprimoramento dos algoritmos
inicialmente desenvolvidos na décadas de 1960 e a1970.

Sobre o uso da análise Fuzzy Cluster em estudos geográficos, Camara et. al. (2003, p. 94) avaliaque a partir da década de 1990 o desenvolvimento da análise de sistemas geográficos deu grande ênfase no uso de técnicas de estatística espacial e lógica nebulosa (fuzzy). Burrough (1986)
destaca da seguinte forma a importância desta técnica para a modelagem dos sistemas geográficos:

> Os limites desenhados em mapas temáticos (como solo, vegetação, o geologia) raramente são precisos e desenha-los como linhas finas muitas vezes não representa adequadamente seu caráter. Assim, talvez não nos devamos preocupar tanto com localizações exatas e representações gráficas elegantes. (Burrough, 1986, apud Camara et. al., 2003).

O uso de algoritmos fuzzy para classificação é comum em estudos ambientais, trabalhos como osde Guerra e Caldas (2011) e Ortega (2001) ilustram possíveis usos de algoritmo fuzzy na classificação do uso e ocupação do solo a partir de imagens digitais. Já em estudos socioeconômicos o uso da classificação fuzzy é mais restrito, e poucos são os trabalhos publicados no Brasil que fazem uso desta técnica. Simões (2003) utiliza a técnica de conjuntonebulosos para identificação de complexos industriais espaciais a partir de uma matriz de acessibilidade espacial. Miranda-Ribeiro e Garcia (2005; 2008) utilizaram a técnica de FuzzyCruster para investigar a relação entre segregação social e segregação espacial em Belo Horizonte, enquanto Nogueira, Garcia e Horta (2011) classificaram os países da América Latinaquanto ao grau de urbanização a partir de uma matriz de particionamento fuzzy. Na mesmalógica dos trabalhos anteriores, Tavares e Porto Junior (2010) utilizaram as técnicas de FuzzyCluster para classificar os municípios da região sul do Brasil de acordo com o nível de desenvolvimento.

Em comum a todos os trabalhos na área socioeconômica acima citados tem o fato de fazerem uso do algoritmo Fanny, implementado no S-PLUS. O fato do pacote estatístico que está entre osmais utilizados nas ciências humanas, o SPSS, não disponibilizar um algoritmo de classificaçãoFuzzy, pode ser um dos fatores a colaborar para o baixo uso desta técnica em estudos socioeconômicos. Assim como a preferência pelo algoritmo Fanny pode ser, em parte, explicadapelo fato de ser este o único algoritmo fuzzy cluster implementado em um pacote estatísticocomercial e de interface gráfica, o S-PLUS. Esta constatação reforça a importância deste trabalho, ao demostrar o uso de outros algoritmos Fuzzy Cluster mediante o uso dos recursosdisponíveis na biblioteca do pacote R.

O R é classificado pela comunidade de usuários como uma linguagem e um ambiente de desenvolvimento integrado, para cálculos estatísticos e gráficos. É distribuído mediante licençaGPL (Licença Pública Geral), e possui uma série de pacotes adicionais com funções específicas.

O uso do R na análise de cluster é facilitado pela disponibilidade na biblioteca do ambiente(CRAN) de uma série de pacotes com diversos algoritmos e técnicas de análise,

permitindo aexploração de diversas metodologias de agrupamento disponíveis na literatura, o que é inviávelnos pacotes estatísticos comerciais.

2 – METODOLOGIA

A análise de agrupamentos, também conhecida como análise de conglomerados, classificação oucluster, tem como objetivo dividir os elementos da amostra, ou população, em grupos de forma que os elementos pertencentes a um grupo sejam similares entre si com respeito às variáveis (características) que neles foram medidas, e os elementos em grupos diferentes sejam

heterogêneos em relação a estas mesmas característica (MINGOTI, 2005, p. 155). Todo processode classificação requer a escolha das unidades observacionais e das variáveis subscritas, que são, na visão de Faissol (1972, p. 78), as duas decisões arbitrárias fundamentais. Estas escolhas refletem o julgamento do investigador sobre os aspectos da realidade que são relevantes para o propósito da classificação desejada, sendo, portanto, uma caracterização inicial do dado, sem direcionamento matemático ou estatístico (Ferreira, Lima, 1979, p. 114). A complexidade do problema de agrupamentos advém da sua natureza não supervisionada, pois não se dispõe de um resultado final desejado (meta concreta a ser alcançada). Assim, a definição do que se entende por grupo (cluster) é carregada de elevado grau de subjetividade, e existe uma variedade de categorizações possíveis para um mesmo conjunto de dados, pois os objetos podem ser agrupados de diferentes maneiras dependendo da perspectiva (Vendramin, 2012).

Fuzzy Cluster é um processo de particionamento não-hierárquico, no qual busca-se a divisão doconjunto de entidades em um número de grupos homogêneos, com relação a uma medida desimilaridade apropriada. Há uma diferença fundamental entre um agrupamento tradicional hard e um agrupamento fuzzy: enquanto no agrupamento hard o objeto pertence a um único cluster, noagrupamento fuzzy um objeto pode pertencer a mais de um grupo, mas com diferentes graus depertinência (u). Portanto, o que define o agrupamento fuzzy é que sua aplicação a uma base dedados com n Objetos X = {X1, ..., Xn}, tem como resultado final uma matriz de partição fuzzydesses objetos em um dado numero de k de grupos, tal que U = Uij kxn, no qual U é uma matriz departição fuzzy k x n cujos elementos Uij representam o grau de pertinência (ou simplesmentepertinência) do j-ésimo objeto ao i-ésimo grupo fuzzy.

Nos algoritmos tradicionais tipo hard assume-se que os objetos pertencem exclusivamente a um único grupo, mas em algumas situações esta hipótese pode não ser realista, pois as distribuiçõesestatísticas possuem sobreposições, ou seja, as categorias se sobrepõem umas às outras (Vendramin, 2012). Os algoritmos de agrupamento difusos foram desenvolvidos para lidarexatamente com este tipo de situação. Atualmente existe uma variedade de algoritmos baseadosem modificações e/ou extensões da formulação original do algoritmo mais conhecido, o Fuzzy c-Means. Esta diversidade de opções à disposição do pesquisador é ilustrada no trabalho deVendramin (2012) que apresenta as principais características e potencialidades do ponto de vistado processamento computacional de um total de 23 diferentes algoritmos para agrupamento fuzzyde dados.

2.1 - Algoritmos selecionados

Para os propósitos deste trabalho foram selecionados apenas três algoritmos para a classificaçãodos municípios brasileiros segundo um conjunto de indicadores de bem estar social: os dois maistradicionais, fanny e c-means, e uma modificação do segundo, o c-medoids. A escolha desteúltimo deve-se a tentativa de explorar a peculiaridade deste algoritmo para a classificação demunicípios, pois, como discutido adiante, os grupos são formados tendo como centro médio nãoos valores médios do grupo segundo as p-variáveis selecionadas, mas um dos elementos dopróprio grupo.

Na operacionalização no R foram utilizados os pacotes cluster para a classificação fanny, fclust para a c-means e vegclust para a c-medoids. O algoritmo c-means também está disponível nospacotes vegclust e e1071, que produziram os mesmo resultados do pacote fclust. Também épossível encontrar o algoritmo c-medoids no pacote fclust, mas neste caso a função disponível produziu resultados inferiores aos encontrados via vegclust, uma análise da função de cálculo doc-medoids no fclust mostrou divergências em relação à notação do artigo de referência.

2.1.1 - Algoritmo FANNY

O método FANNY, desenvolvido nos trabalhos de Bellman, Kalaba e Zadeh (1966) e Ruspini (1969), define que para cada objeto i e cada cluster v, existe uma associação uiv que indica o graude pertencimento do objeto ao cluster. As associações são definidas através de processos iterativos, buscando a minimização da função objetivo J, que segundo notação de Kaufman eRousseeuw (1990) é dada por:

$$J = \sum_{v=1}^{k} \frac{\sum_{i,j=1}^{n} u_{iv}^2 u_{jv}^2 d(i,j)}{2\sum_{j=1}^{n} u_{jv}^2}, \qquad (1)$$

na qual para cada elemento i e cada cluster v há uma pertinência, inicialmente desconhecida, u_{iv}, dada porque indica quão fortemente *i* pertence a *v*, se satisfeitas as condições de $\sum_{i=1}^{N} u_{iv} > 0$ para todo $i=(1,\ldots,n)$, e $\sum_{i=1}^{N} u_{iv} = 1$ para todo $v=(1,\ldots,k)$. Sendo d(i,j) uma medida de dissimilaridade entre os objetos i e j.

2.1.2 - Algoritmo Fuzzy C-Means (FCM)

O algoritmo FCM, desenvolvido por Dunn (1974) e Bezdek (1981, 1984), é, segundo Jain et al. (1999), um dos mais facilmente compreendidos, bem documentados e intensamente utilizadosalgoritmo fuzzy cluster, e consiste numa versão fuzzy do algoritmo de agrupamentos rígido k-Means. No processo de partição do FCM o conjunto X = {x1, x2, ..., xn} é dividido em p grupos,dado o grau de pertinência uij da amostra xi ao j-ésimo grupo na matriz u. Via processos iterativosbusca-se minimizar a função objetivo dada por:

$$J(U;c) = \sum_{i=1}^{n} \sum_{j=1}^{p} u_{ij}^{m}\, d(x_i, c_j)^2 \qquad (2)$$

Sendo $\sum_{i=1}^{n} u_{ij} > 0$ para todo j, e $\sum_{i=1}^{N} u_{ij} = 1$ para todo j, m [1, ∞] é o parâmetro de fuzziness[2], cj é o centro de um agrupamento fuzzy (j=1,...,p) e d(xi,cj) a distância entre xi e cj. Ocentro do agrupamento j é definido como:

$$c_j = \frac{\sum_{i=1}^{n} u_{ij}^{m} x_i}{\sum_{i=1}^{n} u_{ij}^{m}}, \qquad (3)$$

e a função de pertinência fuzzy u é dada por:

[1]Na notação do pacote *cluster* do R (Maechler, 2013), o valor *2* do exponencial da função de pertinência u_{iv},denominado *exponente de pertinência* (*membership exponent*) ou parâmetro de imprecisão (*parameter of fuzziness*), é substituído por uma constante *r,* que tem valor maior que 1 e por padrão valor 2. Aproximando o valor de *r* a 1, gera-se agrupamentos mais rígidos (*crisper clusterings*), com o aumento o valor de *r* em direção ao ∞, aumenta-se o grau de fuzzyness (nebulosidade) do agrupamento, até atingir a completa fuzzyness, no qual todos os elementos pertencem igualmente a cada um dos agrupamentos gerados (quando todo $u_{iv}=1/k$). Nos casos em que pelo menos um elemento tem o valor máximo do grau de pertencimento (u_{iv}) igual em dois ou mais grupos, não é possível a conclusão do processo de classificação, via a desfuzzificação. Nestes casos é necessário reduzir o valor de *r* para aumentar o grau de separação entre os grupos, por isso a notação proposta por Maechler (2013) é mais adequada. Na sistematização feita Kaufman e Rousseeuw (1990, p. 191), os autores reconhecem a possibilidade de variação do valor do exponente *r*, semelhante a opção adotada pelos desenvolvedores do algoritmo *c-means*.

[2]Pal, James e Bezdek (1995, 370) sugerem que a melhor escolha para *m* é, provavelmente, no intervalo [1,25; 2,5], sendo o ponto médio *m* = 2 a escolha preferida dos usuários do FCM.

$$u_{ij} = \frac{\left(\frac{1}{r(x_i,c_j)}\right)^{2/(m-1)}}{\sum_{k=1}^{c}\left(\frac{1}{r(x_i,c_k)}\right)^{2/(m-1)}}. \quad (4)$$

A principal diferença do FCM em relação ao Fanny está na forma de se medir na função objetivo a dissimilaridade intra-grupos. Enquanto o segundo utiliza uma matriz Dxi,xj, ou seja, a distânciaentre os elementos que compõe o grupo, o segundo simplificar o cálculo e faz uso de uma matrizDxicj, calculando apenas a distância entre os elementos do grupo e o centroide do agrupamento.

2.1.3 - Algoritmo Fuzzy C- Medoids - FCMdd

Krishnapuram et al. (1999) e Krishnapuram et al. (2001) criaram a versão Fuzzy do algorírimos de particionamento rígido PAM e CLARA (Kaufman e Rousseeuw, 1990). O algoritmo Fuzzy cmedoids(FCMdd) tem por base uma função objetivo para agrupamento relacional difuso (fuzzy) a partir da identificação de k objetos representativos em cada agrupamento (medoids) queminimizem a dissimiralidade dentro de cada grupo. A principal diferença em relação ao FCM é aseleção como centro do agrupamento não o ponto médio, mas de um elemento dentro do próprioconjunto de dados que minimize a dissimilaridade. O algorítimo foi proposto para aplicação emgrandes bancos de dados voltados ao agrupamento de documentos eletrônicos, mas o princípiode seleção de um dos elemtnos como representativos do Grupo pode ter interessantes aplicaçõesnos estudos socioecnômicos, pois facilita o entendimento e a ilustração das características doGrupo ao selecionar um dos elementos n como síntese das características do Grupo.

Sendo X = {Xi|i=1,⋯,n} um conjunto de n objetos, r(Xi, Xj) expressa a distância entre o objetoXi e o objeto Xj. Enquanto V={v1, v2,⋯,vc}, vi X, representa um subconjunto de X comcardinalidade c, insto é, V é um c-subgrupo de X. Enquanto Xc representa o grupo de todos os subgruposV de X. O algoritmo FCMdd é minimizado pela seguinte função:

$$J_m(V;X) = \sum_{j=1}^{n}\sum_{i=1}^{c} u_{ij}^{m}\, r\left(X_j, V_i\right), \quad (5)$$

onde a minimização é realizada para todo V em X^c. Na função (5) ru_{ij}^{m}senta a matriz depertinência fuzzy de Xj no cluster i, tendo m como exponente de pertinência (parâmetro

defuzziness) com valor no intervalo [1,∞)[3]. A função de pertinência uij é semelhante a utilizada noFCM, dada por:

$$u_{ij} = \frac{\left(\frac{1}{r(x_j, v_i)}\right)^{1/(m-1)}}{\sum_{k=1}^{c}\left(\frac{1}{r(x_j, v_k)}\right)^{1/(m-1)}},$$

onde m [1, ∞] é o parâmetro de fuzziness.

2.1.4 - Medidas de Dissimilaridade

Para agrupar um conjunto de dados constituídos de n elementos com p-variáveis aleatórias em kgrupos, é necessário definir uma medida síntese de dissimilaridade entre os objetos segundo as pvariáveis,de tal forma que quanto menor o valor mais similares/próximos os elementos no planop-dimensional. Nesta situação para cada elemento j tem-se o vetor de medidas Xj, definido por Xj= [X1j, X2j ... Xpj]´, j= 1,2,..., n , onde X1j representa o valor observado da variável 1 medida noelemento j. Existem várias medidas para cálculo da dissimilaridade, e a escolha da medida temimpacto direto no resultado do agrupamento (Mingoti, 2005). Para operacionalização dosalgoritmos selecionados será utilizada a Distância Euclidiana, a "mais popular medida paradados contínuos" (Jain et al., 1999, p. 271, tradução livre), que define a distância entre doiselementos Xi e Xj K, i≠j e K= (1, ..., k), como:

$$d_2(x_1, x_j) = \left(\sum_{k=1}^{d}(x_{i,k} - x_{j,k})^2\right)^{1/2}. \quad (7)$$

2. 2 -Base de dados utilizada

Todo processo de classificação requer "duas decisões arbitrárias fundamentais": a definição dasunidades observacionais e das variáveis subscritas (Faissol, 1972, p. 78). Estas escolhas refletemo julgamento do investigador sobre os aspectos da realidade que são relevantes para o propósitoda classificação desejada, sendo, portanto, uma caracterização inicial do dado, semdirecionamento matemático ou estatístico (Ferreira, Lima, 1979, p. 114).

Neste trabalho propõe-se uma aplicação da técnica fuzzy cluster na classificação dos municípiosbrasileiros segundo indicadores de bem-estar social. Este enunciado já delimita a primeira dasescolhas: as unidades observacionais serão formadas pelo conjunto dos municípios

[3]Os autores recomendam como padrão o valor de *m*=1,5.

brasileiros.Quanto aos indicadores de bem-estar-social a serem utilizados, faz-se necessário uma brevediscussão sobre o conceito em tela.

2.2.1 – Indicadores de bem-estar-social ou de qualidade de vida?

Segundo Ferrán Casas (1999) o debate sobre a questão da qualidade de vida, enquanto conceitooperacionalizável do ponto de vista da pesquisa acadêmica e das intervenções de políticaspúblicas, teve origem nos países desenvolvidos nas décadas de 1960 e 1970, coincidindo com acrise do Estado de bem-estar-social nestes países. Ainda segundo Casas (1999), o avanço dodebate neste período leva ao rompimento com a fundamentação conceitual centrada nascondições materiais de vida (bem-estar-social), sedimentando o entendimento da qualidade devida como impregnada de componentes subjetivos (psicossociais segundo o autor).

Joseph Stiglitz, Amartya Sen e Jean-Paul Fitoussi (2009), em relatório com as conclusões deuma comissão internacional de notáveis sobre como medir o progresso das sociedades,sintetizaram os desafios da mensuração do progresso social em 12 recomendações. Dentre elas asque reforçam a importância dos indicadores de qualidade de vida, que devem incluir indicadoressubjetivos de satisfação individual. Como resumido por Veiga (2010, p. 47), para a Comissão "aqualidade de vida só pode ser medida por um índice composto bem sofisticado, que incorporeaté mesmo as recentes descobertas desse novo ramo que é a economia da felicidade". Trata-sede um reconhecimento explícito das limitações da abordagem objetivista baseada apenas nasescolhas feitas pelos indivíduos e que podem ser observadas, as denominadas preferenciasobservadas (Sen, 1986; IPEA, 2012).

A escolha do termo "indicadores de bem-estar-social" é um cuidado quanto ao uso do termomais popular, qualidade de vida, sem que algumas premissas sejam observadas. Portanto, concorda-se com a definição proposta por Casas (1999, p. 5), para quem o bem-estar-socialrepresenta as condições materiais, objetivamente observadas da qualidade de vida. Enquanto aqualidade de vida é função do entorno material (bem estar social) e do entorno psicossocial (bemestar psicológico).

2.2.2 – Indicadores de bem-estar-social selecionados

Mesmo que não seja o interesse deste artigo explorar as possibilidades e limitações inerentes aoprocesso de escolha de um indicador de bem estar social adequado aos municípios brasileiros,convém uma breve exposição das orientações que embasaram a escolha dos indicadores para oteste aqui proposto.

Parte-se das pistas deixadas por dois importantes estudiosos sobre o desenvolvimento e aurbanização brasileiros. Primeiramente o economista Celso Furtado, um dos maiores estudiososda história social e econômica do Brasil, que em depoimento a uma Comissão do CongressoNacional em 1999 expos da seguinte maneira sua visão sobre a questão da pobreza no Brasil:

> Podemos abordar o problema da pobreza de ângulos diferentes. Três são
> as dimensões que têm preocupado os estudiosos da matéria: 1) a questão
> da fome endêmica, que está presente, em graus diversos, em todo mundo; 2) a questão da habitação popular, que alguns países já encontrousolução; e 3) a questão da insuficiência de escolaridade, que contribuipara perpetuar a pobreza (Furtado, 2002, p. 12).

Recentemente, após a onda de protestos que varreram o Brasil em junho de 2013, a urbanistaErmínia Maricato, em artigo com um sugestivo título "É a questão urbana, estúpido!",manifestou sua visão sobre o fundamental da questão urbana brasileira:

> As cidades são o principal local onde se dá a reprodução da força detrabalho. Nem toda melhoria das condições de vida é acessível commelhores salários ou com melhor distribuição de renda. Boas condiçõesde vida dependem, frequentemente, de políticas públicas urbanas –transporte, moradia, saneamento, educação, saúde, lazer, iluminaçãopública, coleta de lixo, segurança. Ou seja, a cidade não fornece apenas olugar, o suporte ou o chão para essa reprodução social. Suascaracterísticas e até mesmo a forma como se realizam fazem a diferença.(...) A cidade constitui um grande patrimônio construído histórica esocialmente, mas sua apropriação é desigual e o nome do negócio é rendaimobiliária ou localização, pois ela tem um preço devido a seus atributos.(Maricato, 2013, p. 6)

Como mensurar o bem estar social brasileiro considerando a perspectiva apontada pelos autores?Sem dúvida o mais conhecido e utilizado índice para comparação dos municípios brasileiros, oIDH, não é suficiente. É consenso, inclusive entre os próprios formuladores do índice, que o IDHtem seus limites enquanto indicador de progresso social (Sen, 1999). No contexto aqui proposto,o IDH não contempla, pelo menos diretamente, a questão urbana, seja do ponto de vista dascondições de moradia ou do acesso aos serviços urbanos básicos[4]. A variável para o índice derenda, a renda per capita, também não contempla a questão da desigualdade na distribuição intragruposda renda auferida (Gadrey e Jany-Catrice, 2006; Guimarães e Jannuzzi, 2011).

[4]Os indicadores que compõe o IDH-M 2013 são os sete primeiros do Quadro 1.

Entretanto, como não há espaço no âmbito deste trabalho para discussões sobre os indicadoresalternativos, propõe-se partir dos indicadores do IDH e acrescentar três novos indicadores: a taxade pessoas acima da linha de pobreza (inverso da taxa de pobreza); o percentual de domicíliosem condições adequadas de moradia; e o percentual de pessoas que demoram até trinta minutosno deslocamento casa-trabalho (QUADRO 1). O primeiro visa corrigir a deficiência doindicador de renda per capita, apontado os locais onde, embora sejam elevadas as rendas médias,encontra-se elevado índice de pobreza[5]. O segundo visa sintetizar as condições de moradia noBrasil, trata-se de um indicador sugerido pela ONU para monitoramento da meta 11 dosObjetivos de Desenvolvimento do Milênio (ODMs), com metodologia definida para o Brasil noRelatório Nacional de Monitoramento dos ODMs (IPEA, 2007)[6]. Por fim, o terceiro indicadorbusca captar os custos de aglomeração dos grandes centros urbanos, fruto da expansão horizontalda área ocupada e da concentração das atividades geradoras de emprego nas áreas centrais, queimplicam em tempos cada vez maiores de deslocamento no trajeto casa trabalho[7].

[5]O indicador de pobreza exige a definição de uma linha de corte. A linha aqui utilizada considerou o critério doprincipal programa federal de transferência de renda, o Bolsa Família, que em 2010 considerava como pobres, epúblico alvo do Programa, todos aqueles que possuíam renda per capita de até R$140,00. Para uma discussão sobreos resultados para o contexto brasileiro da mensuração da pobreza segundo diferentes linhas de corte ver IPEA(2010). Para um indicador alternativo considerando o patamar médio da renda local via correção da linha por uma
proxy do custo de vida nos municípios brasileiros ver Garcia e Matos (2007).

[6] Na versão do indicador aqui adotado foram feitas pequenas alterações no indicador nacional, conformemetodologia adaptada para o monitoramento local dos ODMs no Município de Belo Horizonte (PBH, 2008).Considera-se como adequados o percentual de domicílios particulares permanentes que atendem simultaneamente
todas as seguintes condições: acesso a rede de esgoto geral ou fossa séptica, acesso a rede geral de água comcanalização em pelo menos um cômodo, coleta de lixo por serviços de limpeza, iluminação via rede geral, existênciade banheiro e média de até 3 moradores por cômodo servindo como dormitório.

[7]Este mesmo indicador, mas um pouco menos restritivo considerando o tempo de deslocamento de até uma hora, foiutilizado no cálculo do índice de bem-estar urbano produzido por pesquisadores do Observatório da Metrópoles, osautores do índice consideraram este indicador como "uma boa proxy das condições de deslocamento, apesar de nãose referir a outros elementos importantes da mobilidade urbana, a exemplo da a sua qualidade, da segurança dosserviços prestados e da infraestrutura disponível" (Ribeiro e Ribeiro, 2012, p. 19). Para exemplificar a importânciados diferenciais entre os municípios brasileiros registrados por este indicador, dados organizados por Matos eFerreira (2013) mostram que se o tempo gasto no deslocamento casa-trabalho fosse remunerado resultaria em umacréscimo médio da remuneração de 22,8% no entorno metropolitano de SP e de 11,5% nas cidades médias, áreasnas quais gasta-se em média, respectivamente, 101 e 51 minutos por dia no deslocamento casa-trabalho-casa.

QUADRO 1 – Indicadores de bem-estar social selecionados

Sigla	Indicador	Fonte
ESPVIDA	Esperança de vida ao nascer	FJP, IPEA, PNUD. Atlas IDH 2012
T_FREQ5a6	Percentual da população de 5 a 6 anos de idade frequentando a escola	
T_FUND11a13	Percentual da população de 11 a 13 anos de idade frequentando os anos finais do fundamental ou que já concluiu o fundamental	
T_FUND15A17	Percentual da população de 15 a 17 anos com fundamental completo	
T_FUND18M	Percentual da população de 18 anos ou mais com fundamental completo	
T_MED18a20	Percentual da população de 18 a 20 anos de idade com o ensino médio completo	
RDPC	Renda per capita média	
PNPOB	Percentual de não pobres (renda per capita maior que R$140,00)	
DOM_ADEQ	Percentual de Domicílios com Condições Adequadas de Moradia	IBGE. Censo 2010 Microdados da Amostra (elaboração própria)
DESL_ATE30	Percentual de pessoas que gastam até 30 minutos no deslocamento casa trabalho	

2.3 - Quantos grupos?

Um dos principais problemas na análise de cluster é a definição do melhor algoritmo e donúmero total de cluster k a utilizar. Embora nos métodos não-hierárquicos o número de clusterseja pré-definido, é necessário avaliar em que medida o total de agrupamentos selecionados seadequa ao agrupamento natural das variáveis selecionadas no espaço p-dimensional. Não existeuma resposta exata para esta questão, e na ausência de critérios estatísticos internos usado parainferência, encontra-se na literatura uma série de critérios ad hoc que podem auxiliar na decisãofinal (Mingoti, 2005; Hair Jr. Et al., 2005).

Uma abordagem utilizada para contornar este problema consiste em executar o algoritmo deagrupamento diversas vezes com matrizes de distância e de partições/protótipos iniciaisdiferentes e número de grupos variados e então escolher a partição mais adequada de acordo comum critério[8]. É possível utilizar diversas medidas como regra de parada (Stopping Rule) que

[8]G. Milligan e M. Cooper testaram um total de 30 medidas de validação de agrupamento visando determinar onúmero ideal de cluster, denominadas *stopping rule* (regra de parada), no intervalo [2, 5] para cada uma das medidas utilizadas em um processo de agrupamento hierárquico Os resultados apontaram para os métodos de validaçãoexterna Calinski and Harabasz Index (PseudoF) e Duda and Hart index (Pseudo T2) como bom indicadores do

possibilitam a identificação do numero ótimo de grupos a serem definidos para o conjunto dedados. A Figura 1 apresenta a aplicação no banco de dados utilizado, via a função clValid dopacote clValid disponível no R, de duas medidas para avaliar o grau de separação das partições:o Silhouette Width e o Dunn Index[9]. As medidas foram computadas para os resultados dosagrupamentos gerados por 10 diferentes métodos no intervalo de 2 a 8 partições geradas. Comoambos os índices devem ser maximizados, os resultados mostram melhores resultados doSilhouette Width para partições com menor número de grupos, enquanto para o Dunn Index, nogeral, partições com maior número de grupos geraram melhores resultados, embora como maiorvariação entre os diversos algoritmos utilizados.

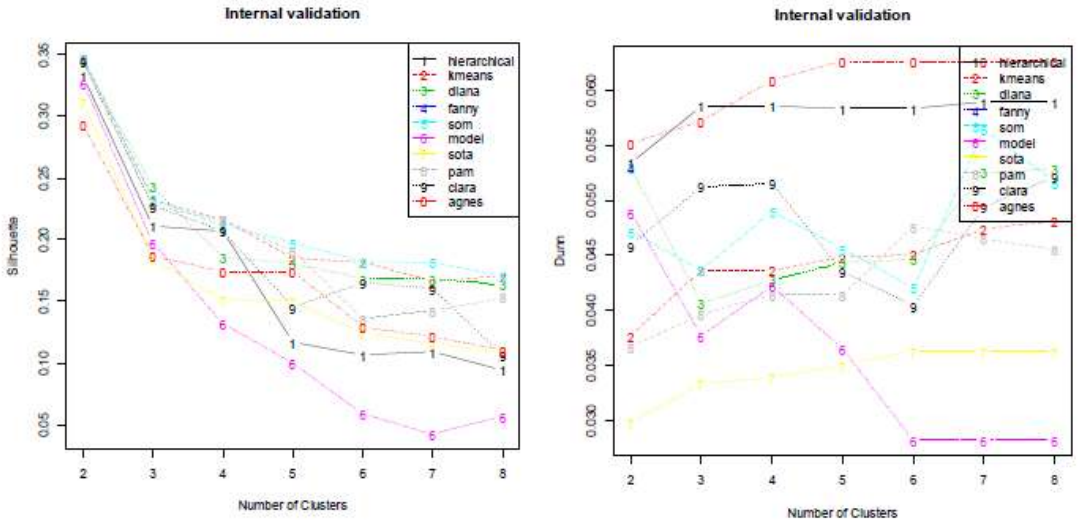

Figura 1: estimativa dos coeficientes de silhueta e de Dunn segundo valores de k no intervalo [2-8] para dez diferentes algoritmos.

Optou-se por utilizar uma classificação em três grupos. Considerando o tipo de observaçãoutilizada no agrupamento, a geração de três grupos de municípios ilustra melhor as diferenças donível de desenvolvimento do municípios brasileiros, permitindo separar dois grupos com osmelhores e os piores indicadores em dois polos, separados por um grupo intermediário, detransição.

3 - RESULTADOS

número de Grupos (Milligan e Cooper, 1985, p. 169; Mingotti, 2005).

[9]A *Silhouette Width* é a largura média da silhueta das observações (ver detalhamento no item 3, função (8)). O*Dunn Index* é a razão entre a menor distância entre as observações que não estão no mesmo cluster e a maiordistância entre as observações dentro do mesmo clusters, dada por $D = \min_{i \in I_k, j \in I_{k'}} \left\| M_i^{\{k\}} - M_j^{\{k'\}} \right\| / \max_{i,j \in I_k, i \neq j} \left\| M_i^{\{k\}} - M_j^{\{k\}} \right\|$.
Tem valor entre 0 e o infinito, e deve sermaximizada (Brock et al., 2013).

Os resultados gerados pelo agrupamento podem ser visualizados nos gráficos da figura 2, queplotam o agrupamento num plano bidimensional tendo com eixos dois componentes que explicam 68% da variabilidade original da matriz de variáveis. Observa-se que o desenho doagrupamento intermediário foi o quem mais sofreu alterações entre os três algoritmos[10].

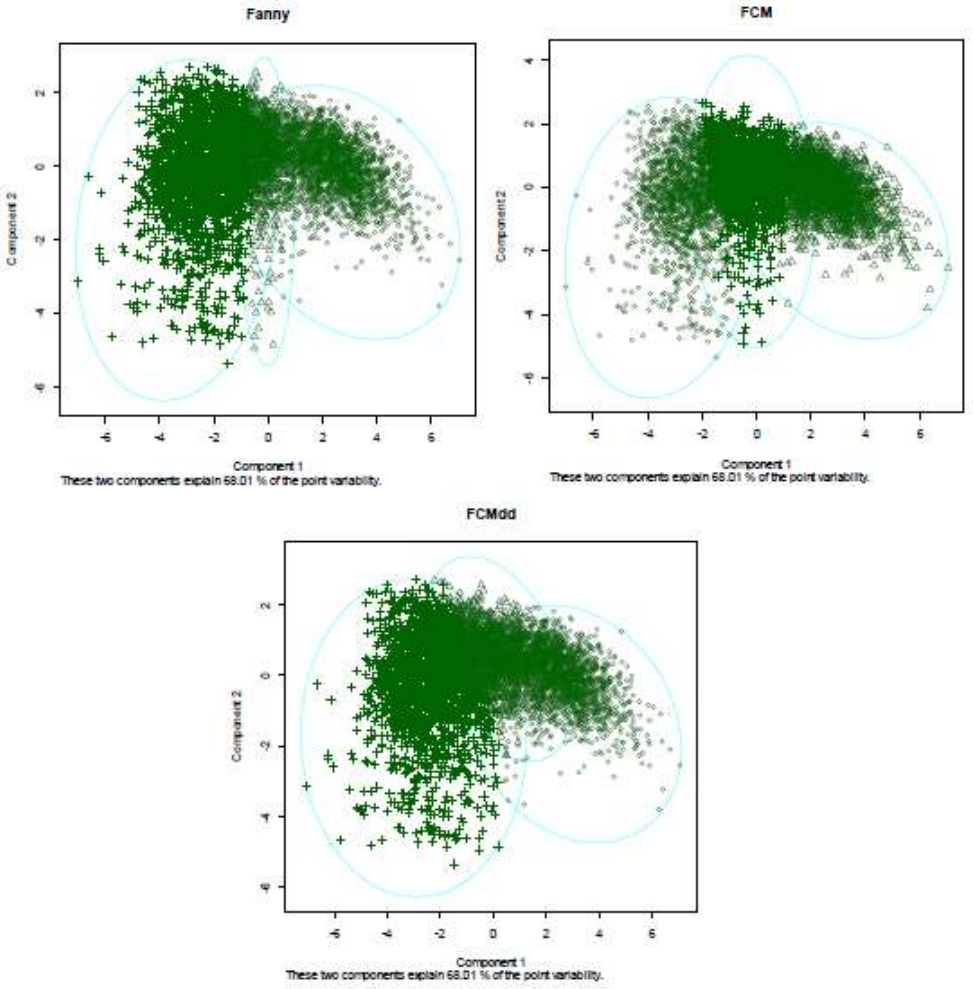

Figura 2: Gráficos Clusplot para resultados da classificação gerada pelo algoritmosselecionados.

O comportamento espacial das mudanças no grupo intermediário (cluster 2 no Fanny e FCMdd e3 no FCM) pode ser observado na figura 3, que apresenta os resultados da classificação dosmunicípios brasileiros segundo as três técnicas utilizadas.

[10]Foi utilizado o valor do parâmetro de fuzziness *m*=1,9 em todos os algoritmos. O valor padrão 2 não foi utilizadpois no algoritmo fanny gerava casos de fuzzyness.

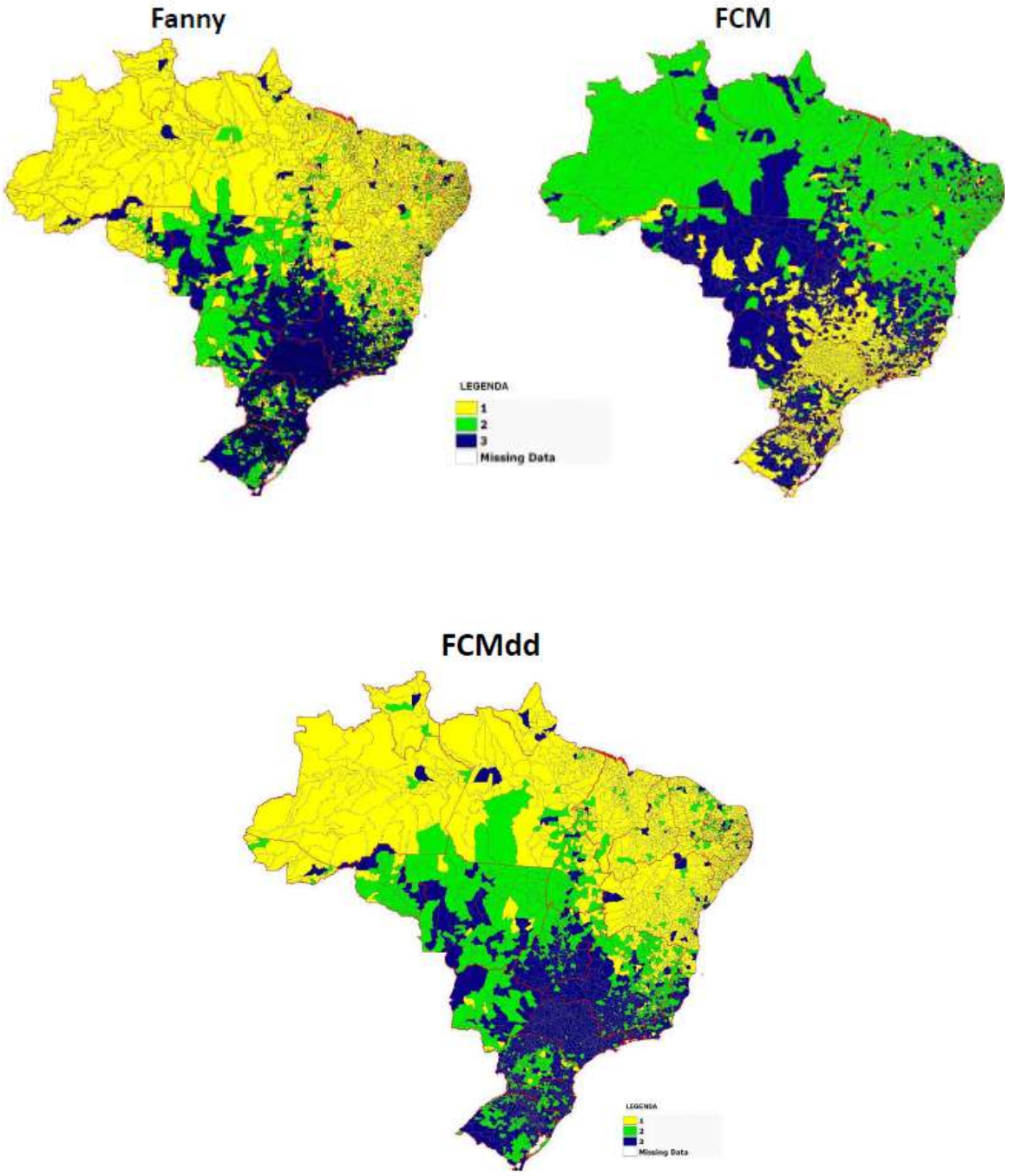

Figura 3: classificação dos municípios brasileiros quanto ao cluster de pertencimentosegundo os algoritmos utilizados.

A classificação apresentada na Figura 1 pode ser interpretada a luz dos valores médios(centroides) de cada agrupamento para as varáveis selecionadas (Tabela 1). O FCMdd selecionou

como centroides (medoids) os municípios de Cravolândia-BA (cluster 1), Presidente Kennedy-ES (cluster 2) e Piraí do Sul-RS (cluster 3).

Tabela 1: valores médios das variáveis (centroides) dos agrupamentos gerados pelos algoritmos selecionados.

	Fanny			FCM			FCMdd		
Variável	C1	C2	C3	C1	C2	C3	C1	C2	C3
ESPVIDA	70,7	73,8	75,3	75,3	70,4	73,7	70,9	73,5	73,9
T_FREQ5A6	90,8	87,8	93,8	94,0	92,0	90,1	93,1	89,8	91,0
T_FUND11A13	78,8	87,2	90,1	89,9	78,0	87,2	82,4	87,4	89,5
T_FUND15A17	41,6	56,9	67,2	67,2	40,1	57,9	42,0	55,8	62,5
T_FUND18M	31,8	38,4	47,9	48,9	31,2	39,6	33,8	36,9	46,1
T_MED18A20	23,9	35,8	48,4	48,5	23,0	37,7	27,9	38,0	43,6
RDPC	281,4	483,0	709,5	715,5	274,7	509,6	259,9	420,0	631,8
PNPOB	59,1	82,2	92,8	92,3	57,6	81,5	57,5	79,5	88,1
DOM_ADEQ	13,2	24,3	52,2	57,8	12,8	27,1	18,8	27,9	47,0
DESL_ATE30	80,6	82,0	80,4	80,5	80,1	82,3	78,2	85,1	78,4

Fonte: citadas no Quadro 1

Outra forma de visualizar graficamente o resultado do particionamento é através do gráfico desilhueta (Rousseeuw, 1987) com a medição da largura média da silhueta das observações dadapor s(j). No processo de cálculo de s(j) primeiro é obtido o valor individual de silhueta para cadaobservação s(i), dado pela dissimilaridade média do objeto i para os demais objetos do Grupo A,indicado como a(i), e a dissimilaridade média do objeto i para os objetos do grupo B, d(i,B).Após computar d(i,B) para todos os grupos B≠A, o menor valor é selecionado, isto é, b(i) = mind(i,B), B ≠ A, que representa a dissimilaridade média do de i ao grupo mais próximo. Assim ovalor s(i) é dado por:

$$s(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}},$$

sendo que a média dos valores de s(i), dada $s(j) = \sum_{i=1}^{n} s(i)/n,$ confiança do agrupamento, com valores no intervalo [-1,1], quanto mais próximo de 1 melhor aqualidade do agrupamento e quanto mais próximo de -1 pior.

Os resultados de s(i) e s(j) apresentados na Figura 4 mostram melhores resultados de s(j) para ométodo FCM (s(j)= 0,228), no qual apenas no cluster 1 é possível encontrar objetos com valoresnegativos de s(i), que indica objetos que, na média, estão mais próximos de objetos externos aogrupo de classificação. Os algoritmos Fanny e FCMdd obtiveram, respectivamente, valores des(j)= 0,174 e s(j)= 0,208. Outra observação interessante é a capacidade do FCM gerar grupos detamanhos semelhantes, o que pode ser uma característica interessante do algoritmo

para aclassificação em questão se esta vier acompanhada de bons resultados na análise de validação.

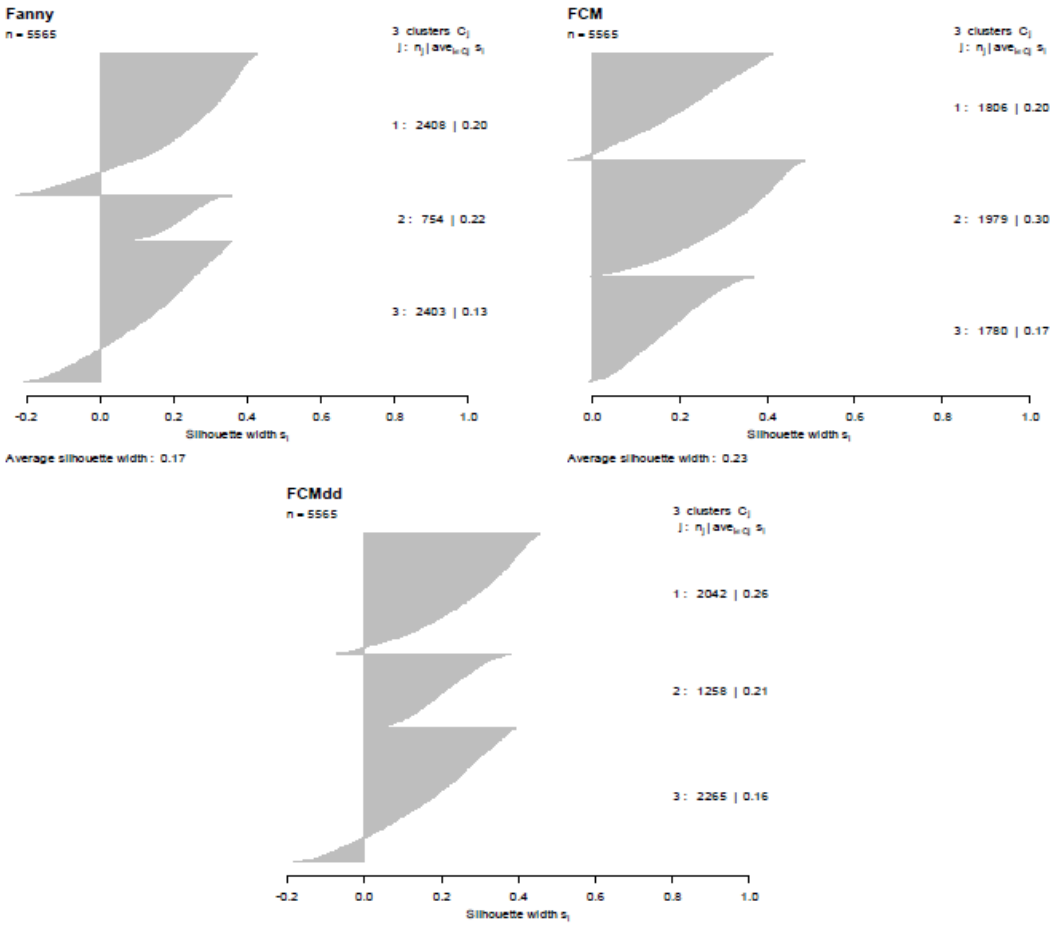

Figura 4: Gráfico de Silhueta para resultados da classificação gerada pelo algoritmosselecionados.

3.1 - Validação dos Agrupamentos

O objetivo de uma partição não-hierarquica é sempre produzir uma partição dos n elementos emk grupos de modo que a partição satisfaça dois requisitos básicos: coesão interna e isolamentodos grupos formados (Mingoti, 2005, p. 192). Segundo Mingoti (2005), quando os dadosapresentam uma partição natural no espaço em relação as p-variáveis selecionadas, os resultadosdo agrupamento não serão muito diferentes entre os diversos métodos. Mas quando a partiçãonatural apresenta grande interseção entre grupos, os resultados serão diferentes de acordo com ométodo e parâmetros escolhidos. Neste contexto o termo validação é utilizado para caracterizar,de forma ampla, os diferentes procedimentos para avaliar de maneira objetiva e quantitativa osresultados da análise de agrupamento. Os índices e critérios podem ser de três tipos (Vendraminet al., 2010; Campello, 2010; Desgraupes, 2013):

a. Externos: Avalia o grau de correspondência entre a estrutura de grupos (partição ou hierarquia) sob avaliação e informação a priori na forma de uma solução de agrupamento

esperada ou conhecida ou mesmo uma outra partição para comparação.

b. Internos: Avalia o grau de compatibilidade entre a estrutura de grupos sob avaliação e osdados, usando apenas os próprios dados

c. Relativos: Avaliam qual dentre duas ou mais estruturas de grupos é melhor sob algum

aspecto. É uma classe particular de critérios com habilidade para indicar qual a melhor dentre duas ou mais partições.

O Objetivo é sempre avaliar individualmente uma única partição e quantificar esta avaliaçãoatravés de um valor que possa ser comparado relativamente. No clássico trabalho de Milligan eCooper (1985) os autores compararam 30 diferentes medidas de validação de agrupamentosatravés de um extenso conjunto de experimentos envolvendo vários conjuntos de dados.Vendramin et al. (2010) testou 40 critérios de validação relativos com foco na complexidadecomputacional requerida no processamento. Desgraupes (2013) sistematizou e disponibilizou nopacote clusterCrit para R um total de 27 índices de validação interna ou relativos e outros 11 devalidação externa. A existência na literatura de dezenas de tais critérios torna uma tarefa difícilpara o usuário a escolha de um critério específico, e a alternativa sugerida tem sido o usocombinado de diversos critérios como forma de avaliar o agrupamento ideal.

A maioria dos índices referidos acima não fazem distinção entre o tipo de agrupamento hard oufuzzy, pois não consideram em seus cálculos a matriz de pertinência U. Embora em menornúmero, também existem diversos critérios para avaliação de partições fuzzy (Campello, 2010).Vendramin (2012) apresenta uma proposta de agregação dos principais critérios validaçãorelativa de agrupamentos fuzzy em dois grandes grupos[11]:

a. Índices baseados apenas na matriz de partição: avaliam a qualidade das soluções utilizando as informações apenas da matriz de partição U. Não utilizam informações sobre os atributos dos objetos e das dissimilaridade entre eles. Ex: Partition Coeficient (PC), Partition Entropy (PE) e Modified Partition Coefficient (MPC).

[11]Embora Vendramin (2012) detalhe um total de 13 índices de validação de agrupamento fuzzy, optou porexemplificar em cada categoria apenas os índices a serem utilizados para análise subsequente. A escolha destesíndices baseou-se em dois critérios: primeiro a disponibilidade de funções nos pacotes R utilizados que permitissemo cálculo, e segundo a viabilidade de uso das funções disponíveis para cálculo do índice na comparação dediferentes algoritmos. Algumas funções, da forma como foram implementadas nos pacotes R, permitem apenas aavaliação de resultados FCM.

b. Índices Baseados na Matriz de Dados: usam tanto a matriz de partição quanto a matrizde dados. EX: Xie-Beni (XB), Fuzzy Simplified Silhuette (FSS).

Para os testes de validação aqui propostos foram selecionados cinco índices disponíveis viafunção fclust.index do pacote fclust no R. O Quadro 2 detalha as fórmulas de cálculo e asreferências de cada um dos índices utilizados. As seguintes notações são utilizadas no Quadro 2:

N é número de elementos, Uij os valores da matriz de pertencimento, vj o centro do cluster e k onúmero de clusters.

QUADRO 2 - Fórmulas de cálculo e referências dos índices de validação de agrupamentos fuzzy utilizados

Nome	Fórmula	Referência Bibliográfica
PC (Partition Coeficient)	$V_{PC} = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{N}(u_{ij})^2$	Bezdek, 1981; Dunn, 1974
MPC (Modified Partition Coefficient)	$V_{MPC} = 1 - \frac{k}{k-1}(1 - V_{PC})$	Dunn, 1974; Dave, 1996
PE (Partition Entropy)	$V_{PE} = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{N} u_{ij} log_a(u_{ij})$	Bezdek, 1981
Xie Beni Index	$V_{XB} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{N}(u_{ij})^m \|x_j - v_i\|^2}{N \min_{l \neq s}\|v_l - v_s\|^2}$	Xie & Beni, 1991;
FSS (Fuzzy Simplified Silhouette index)	$V_{FSS} = \frac{\sum_{j=1}^{N}(u_{pj} - u_{qj})^{\propto} S_j}{\sum_{j=1}^{N}(u_{pj} - u_{qj})^{\propto}}$ Sendo S_j a silhueta individual do objeto X_j conforme definição da função (8).	Campello & Hruschka, 2006

Fonte: adaptado de Vendramin (2012), Giordani e Ferraro (2013) e Meyer (2013)

O Quadro 3 apresenta os resultados dos índices selecionados. Observa-se que nos cinco critériosutilizados o FCM apresentou melhores resultados, seguido pelo FCMdd que não superou oFanny apenas no Xie Beni Index.

QUADRO 3 – Resultados dos índices de validação de agrupamento fuzzy selecionados

Nome	Objetivo	Fanny	FCM	FCMdd
PC (Partition Coeficient)	max	0,353	0,519	0,372
MPC (Modified Partition Coefficient)	max	0,030	0,279	0,058
PE (Partition Entropy)*	min	1,069	0,818	1,041
Xie Beni Index**	min	0,806	0,680	1,188
FSS (FuzzySimplified Silhouette Index) ***	max	0,271	0,370	0,333

Notas sobre as adaptações realizadas:
* utilizou-se exponencial de 1 como base do logaritmo. Para o FCMdd não é possível calcular o termo YZ[lQm ____ quando Nj _ V, ou seja, para os elementos tomados como centroides do agrupamento, que possuem *u*=1 para o cluster de centro e *u*=0 para os demais. Os elementos centroides de grupos foram excluídos do cálculo e utilizou-se N-k.
** a função objetivo fanny não utiliza a distância em relação ao centro médio do grupo, por isso foi necessário calcular os centros médios dos agrupamento gerados por este método.
*** utilizou-se α =1

3.1.1 - Análise discriminante para validação de agrupamento

Como critério adicional para validação do agrupamento é possível implementar uma AnáliseDiscriminante (AD) a partir do agrupamento gerado pela análise de cluster. Segundo Mingoti(2005) a AD é uma técnica que pode ser utilizada para classificação de elementos, mas requer apredefinição dos grupos, e de suas características, para os quais cada elemento pode ser classificado. O que permite a elaboração de uma função matemática, chamada regra declassificação ou de discriminação, que pode ser utilizada para classificação dos novos elementos.

Aqui propõe-se o uso da técnica de análise discriminante para mensuração do grau de ajuste daclassificação, mas sem entrar em maiores detalhes sobre esta técnica de análise que o objetivodeste artigo. A técnica de validação adotada consiste em dividir o grupo previamente classificadopelos algoritmos selecionados em dois subgrupos, um para treinamento e elaboração da funçãode discriminação, e outro para aplicação da função elaborada para o primeiro grupo.

De forma semelhante à análise de cluster, na análise discriminante pretende-se explorar os dadosde forma a maximizar a diferença entre os K grupos existentes. Para isso são criadas funçõespreditivas, combinações lineares das variáveis preditivas, que discriminam os K grupos maximizando a diferença entre eles. Assim, dado n elementos, K grupos e p variáveis preditivas X=(X1,...,X p), o objetivo é estabelecer uma combinação linear dada por $Y_j = v_0 + \sum_{i=1}^{p} v_{ji} X_i$, onde Y é o vetor das funções lineares Yj, v a matriz dos coeficientes lineares vji e X o vector das
variáveis preditivas (Mingoti, 2005).

Os resultados oferecem mais uma medida para avaliar o grau de confiabilidade, ou qualidade, doesquema classificatório da análise de cluster. A análise discriminante, quando comparada aosíndices de validação apresentados no tópico anterior, oferece como vantagem o uso de umamedida de validação que vai além do uso apenas das distâncias matemáticas, inserindo uma regrade classificação fundamentada na teoria das probabilidades, com o objetivo de básico de derivarcombinações lineares das p variáveis iniciais que maximizem a

diferenciação entre os grupos. Atécnica tem como principais pressupostos a normalidade das variáveis, a homogeneidade dasmatrizes de variância-covariância e a existência de diferenças significativas entre os grupos(Mingoti, 2005; Ferreira e Lima, 1978).

Os resultados da AD aplicada aos três agrupamentos gerados são apresentados no Quadro 4,enquanto a Figura 5 ilustra a separação dos grupos pela primeira função discriminante. Osresultados da avaliação da reprodutibilidade do modelo de agrupamento para três clusters em quea previsão para o grupo é a classificação prevista pelas funções discriminantes, mostram que,embora os três algoritmos tenham permitido a geração de modelos com alto percentual devariância explicada pela primeira função, observa-se um maior grau de acerto para as funçõesgeradas a partir do FCM (94,4%), confirmando os resultados das técnicas de validaçãoapresentadas no item anterior.

Quadro 4: Síntese dos resultados da análise discriminante para os agrupamentos gerados pelos algoritmos selecionados

Medida	FANNY	FCM	FCMdd
% da variância explicada pela função 1	92.3%	91.9%	90.8%
% de classificação correta	85.6%	94.4%	91.9%

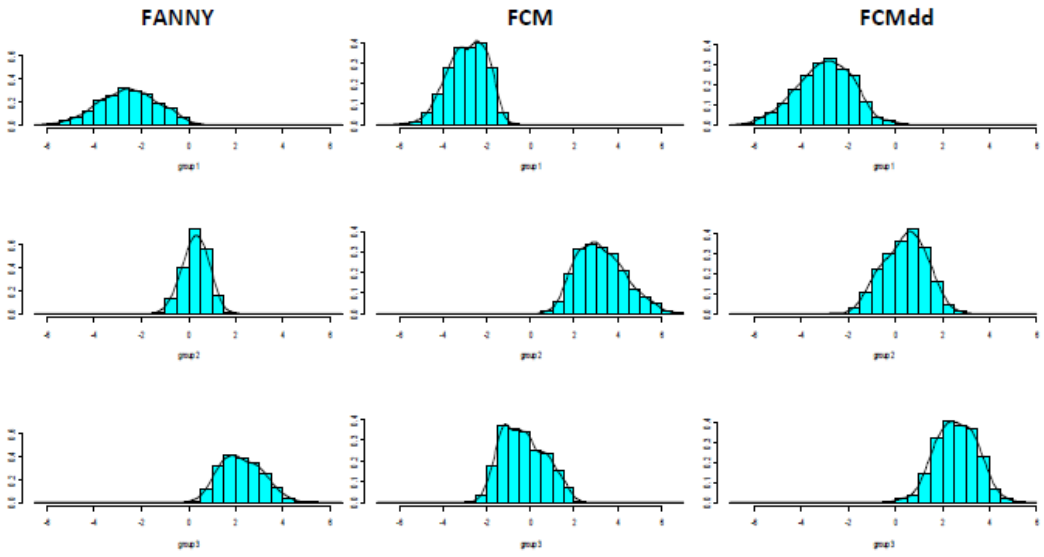

Figura 5: Histogramas da distribuição dos elementos dos grupos segundo scores da função discriminante 1 por algoritmo utilizado.

A AD também permite identificar as variáveis que mais contribuem para a discriminação dosgrupos, através do índice de correlação linear das variáveis originais com os scores da função 1,conforme apresentado no Quadro 5. Observa-se que as variáveis _FREQ5A6 e DESL_ATE30

foram as que apresentaram os menores índices de correlação com a função discriminante demaior poder explicativo do agrupamento, um indicativo de a inclusão do indicador de tempo de

deslocamento pouco contribui com os resultados do agrupamento gerado.

Quadro 5: Coeficiente de correlação linear de Pearson entre as variáveis e os *scores* da função discriminante 1 por algoritmo utilizado

Variável	FANNY	FCM	FCMdd
ESPVIDA	0.877	-0.875	0.884
T_FREQ5A6	0.169	-0.155	0.095
T_FUND11A13	0.664	-0.672	0.666
T_FUND15A17	0.844	-0.843	0.833
T_FUND18M	0.765	-0.792	0.767
T_MED18A20	0.853	-0.845	0.825
RDPC	0.892	-0.885	0.876
PNPOB	0.965	-0.952	0.969
DOM_ADEQ	0.675	-0.712	0.667
DESL_ATE30	-0.002	0.028	-0.027

4 - CONCLUSÕES

Os resultados da análise fuzzy cluster, através do uso de três algoritmos fanny, c-means e cmedoidspara classificação dos municípios brasileiros segundo indicadores de bem-estar-social,mostraram melhores resultados do algoritmo FCM. O FCM obteve melhor desempenho em todasas medidas utilizadas para validação de agrupamento difuso quanto a coesão interna e aoisolamento dos grupos formados. O FCM também obteve bom desempenho nos resultados daavaliação ad hoc via análise discriminante, que foi capaz de reproduzir a classificação gerada pelo FCM com 94,5% de classificação correta. Como este algoritmo não está implementado nos

pacotes estatísticos mais utilizados estudos na área social, o SPSS e o S-PLUS, o trabalho aquiapresentado destaca a importância do uso do ambiente R. Que além de ampliar o leque dealgoritmos disponíveis para a classificação fuzzy cluster, permite ao usuário acessar uma série de

medias de validação de agrupamento e assim explorar toda a potencialidade desta técnica declassificação.

REFERÊNCIAS BIBLIOGRÁFICAS

Bellman , R., Kalaba, R., and Zadeh , L.A., (1966) Abstraction and pattern classification, I. Math. Anal.and Appl. 2, 581-586.

Bezdek, J. (1981). Pattern Recognition with Fuzzy Objective Function Algorithms. Kluwer AcademicPublishers, Norwell, MA, USA.

Bezdek, J. C. (1981). Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press.

Bezdek, James C.; Ehrlich, R.; Full, W. FCM: The fuzzy c-means clustering algorithm, Computers &Geosciences, Volume 10, Issues 2–3, 1984, Pages 191-203

Brock, G., Pihur, V., Datta , S., Datta, S, Package 'clValid'. August 29, 2013, http://cran.rproject.
org/web/packages/clValid/index.html

Brock, G., Pihur, V., Datta, S. and Datta, S. (2008) clValid: An R Package for Cluster Validation Journalof Statistical Software 25(4) http://www.jstatsoft.org/v25/i04

BURROUGH, P. Principles of Geographical Information Systems for Land Resources Assessment.Oxford, England, Oxford University Press, 1986.

CÂMARA, G.; MONTEIRO, Antônio Miguel; MEDEIROS, José Simeão de. Representações Computacionais do Espaço: Fundamentos Epistemológicos da Ciência da Geoinformação. Geografia(Rio Claro), Rio Claro, Brasil, v. 28, n.1, p. 83-96, 2003.

Campello, R. & E. Hruschka (2006). A fuzzy extension of the silhouette width criterion for clusteranalysis. Fuzzy Sets and Systems 157 (21), 2858 - 2875.

Campello, R. J. G. B. Generalized External Indexes for Comparing Data Partitions with OverlappingCategories. Pattern Recognition Letters, Vol. 31, p. 966-975, 2010

CRUZ, A. J. de Oliveira. Lógica Nebulosa. UFRJ, 17 de junho, 2004
Dave, R. N. (1996). Validating fuzzy partitions obtained through c-shells clustering. Pattern RecognitionLetters 17 (6), 613 - 623.

Desgraupes, Bernard. An R Package for Computing Clustering Quality Indices. University Paris Ouest,Lab Modal'X. April 2013, Vignettes do pacoteclusterCritdisponívelem: http://cran.rproject.org/web/packages/clusterCrit/index.html

Dunn, J. (1974). A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. Journal of Cybernetics, 3:32–57.

Faissol, Speridião. "Análise Fatorial: problemas e aplicações na geografia, especialmente nos estudosurbanos." Revista Brasileira de Geografia 34.4 (1972): 7-100.

Ferrán Casas. Calidad de vida y calidad humana. Papeles del Psicólogo, Noviembre , nº 74 , 1999,
disponível in: http://www.papelesdelpsicologo.es/vernumero.asp?id=812

FERREIRA, M. L; LIMA, O.M.B. Processo de Classificação. in: Faissol, Speridião. Tendências atuais nageografia urbano/regional: teorização e quantificação. Fundação Instituto Brasileiro de Geografia eEstatistica, 1978.

Furtado, Celso. Em Busca de Novo Modelo. São Paulo: Editora Paz e Terra, 2002.

GADREY, J.; JANY-CATRICE, F. Os novos indicadores de riqueza. São Paulo: Senac, 2006.

GARCIA, R. A.; MATOS, R. E. S. Mensurando a inserção social dos migrantes brasileiros. In Anais doV encontro nacional sobre migrações, Campinas, 2007.

Gath, I.; Geva, A. B. Unsupervised Optimal Fuzzy Clustering, IEEE Transactions on Pattern Analysis andMachine Intelligence, vol. 11, n. 7, p. 773-781, 1989.

Giordani, P, Ferraro, M. B. Package 'fclust', August 29, 2013, disponível em: http://cran.rproject.
org/web/packages/fclust/index.html

GUIMARÃES, José Ribeiro Soares; JANNUZZI, Paulo M. IDH, Indicadores Sintéticos e suas aplicaçõesem políticas públicas: uma análise crítica. Revista Brasileira de Estudos Urbanos e Regionais, v. 7, n. 1, p.73-90, 2011.

Hair Jr., J.F., Anderson, R.E., Tatham, R. L., Black, W.C. Análise multivariada de dados. trad. AdonaiSchlup Sant'Anna e Anselmo Chaves Neto.- 5. ed.- Porto Alegre: Bookman, 2005.

IPEA - - Instituto de Pesquisa Econômica Aplicada. Objetivos de desenvolvimento do milênio: relatórionacional de acompanhamento. Coordenação: IPEA e Secretaria de Planejamento e InvestimentosEstratégicos. Supervisão: Grupo técnico para o acompanhamento dos ODM. Brasília: Ipea, MP, SPI,2007. 152p

IPEA - Instituto de Pesquisa Econômica Aplicada. PNAD 2009 – Primeiras Análises: Distribuição deRenda entre 1995 e 2009. Comunicados do Ipea Nº 63, 05 de outubro de 2010, Disponível
em:http://www.ipea.gov.br/portal/images/stories/PDFs/comunicado/101005_comunicadoipea63.pdf

IPEA - Instituto de Pesquisa Econômica Aplicada. 2012: Desenvolvimento Inclusivo Sustentável?Comunicados do Ipea Nº 158, 18 de dezembro de 2012, Disponível em: http://www.ipea.gov.br/portal/images/stories/PDFs/comunicado/121218_comunicadoipea158.pdf

Jain, A. K., Murty, M. N., and Flynn, P. J. (1999). Data Clustering: a Review. ACM Comput. Surv.,31(3):264–323.

JANÉ, D. de Almeida. Uma introdução ao estudo da lógica Fuzzy. Hórus – Revista de Humanidades eCiências Sociais Aplicadas, Ourinhos/SP, no. 02, 2004.

Kaufman, L. and Rousseeuw, P.J. (1990). Finding Groups in Data: An Introduction to Cluster Analysis.New York: John Wiley & Sons, Inc.

Krishnapuram R., Joshi A., Nasraoui O., Yi L., 2001. Low-complexity fuzzy relational clustering
algorithms for web mining. IEEE Transactions on Fuzzy Systems, 9, 595-607.

Krishnapuram, R., Joshi, A., & Yi, L. (1999). A Fuzzy relative of the k-medoids algorithm with application to web document and snippet clustering. IEEE International Fuzzy Systems (pp.1281–1286).

Maechler, Martin. Package 'cluster', Version 1.14.4, August 29, 2013, disponívelem:http://cran.rproject.org/web/packages/cluster/index.html
MARICATO, Ermínia. É a questão urbana, estúpido!. Le Monde Diplomatique Brasil, ano 7, n. 73, p. 6-7, agosto de 2013.

MATOS, Ralfo; FERREIRA, Rodrigo Nunes. Espaço, população e economia; dos subespaços proeminentes ao transporte público premente. Anais do II CONINTER - Congresso InternacionalInterdisciplinar em Sociais e Humanidades. Belo Horizonte, de 8 a 11 de outubro de 2013.

Meyer, David. Package 'e1071'. August 29, 2013, disponível em: http://cran.rproject. org/web/packages/e1071/index.html

Milligan, G. W.; Cooper, M. C. An examination of procedures for determining the number of clusters in adata set, Psychometrika 50(2) (1985), 159–179.

Mingoti, S.A. Análise de dados através de métodos de estatística multivariada – uma abordagem aplicada. Belo Horizonte: Editora: UFMG, 2005. 295p.

MIRANDA-RIBEIRO, A. ; GARCIA, R. A. . Segregação socioespacial em Belo Horizonte: uma
aplicação dos modelos difusos. Geografias (UFMG), v. 1, p. 86-97, 2005.

NOGUEIRA, Marly; GARCIA, Ricardo Alexandrino; HORTA, Célio Augusto da Cunha; MOREIRA,Kely Cristina. A urbanização na América Latina e Caribe: um indicador de âmbito regional. RevistaGeográfica de América Central, Universidad de Costa Rica, 2011. p. 1-18.

OLIVEIRA, Suzana Abreu de Souza. Alguns comentários sobre a teoria Fuzzy. Exacta, n. 1, abril, 2003,p. 139-147,Universidade Nove de Julho.

Pal, Nikhil R., and James C. Bezdek. "On cluster validity for the fuzzy c-means model." Fuzzy Systems,IEEE Transactions on 3.3 (1995): 370-379.

PBH - Prefeitura de Belo Horizonte (MG). Relatório de acompanhamento dos Objetivos de Desenvolvimento do Milênio Belo Horizonte – 2008. Secretaria Municipal de Planejamento, Orçamento eInformação -SMPL. Belo Horizonte: SMPL, 2008

Ribeiro, Luiz Cesar de Queiroz, Ribeiro, Marcelo Gomes (Orgs). IBEU: índice de bem-estar urbano. 1.ed. - Rio de Janeiro: Letra Capital, 2013.

Rousseeuw, Peter J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis.Journal of computational and applied mathematics 20 (1987): 53-65.

Ruspini , E.H., (1969) A new approach to clustering, Inform. and Control 15, 22-32.Sen, Amartya. (1995) Rationality and social choice. American Economic Review 85 (1).

Sen, Amartya. Evaluación del Desarrollo Humano Contrbuicion Especial. PNUD - Programa de lasNaciones Unidas para el Desarrollo. Informe sobre Desarrollo Humano 1999. Madrid: Ediciones Mundi-Prensa, 1999.

SIMÕES, R. F. Complexos industriais no espaço: uma análise de fuzzy cluster. TD 209. Belo Horizonte:UFMG/Cedeplar, 2003.

STIGLITZ, J.; SEN, A e FITOUSI, Report by the Commission on the Measurement of EconomicPerformance and Social Progress, 2009. September, 2009.
VEIGA, José Eli da. Indicadores de sustentabilidade. Estudos Avançados, São Paulo, v. 24, n. 68, 2010.

Vendramin, L., Campello, R. J. G. B. &Hruschka, E. R. "Relative Clustering Validity Criteria: A Comparative Overview" Statistical Analysis and Data Mining, Vol. 3, p. 209-235, 2010

Vendramin, Lucas. Estudo e desenvolvimento de algorítimos para agrupamento fuzzy de dados emcenários centralizados e distribuídos. Dissertação (MEtrado - Programa de Pós-GRaduação em Ciênciasde Computação e Matemática Computacional) Institutu de Ciências Matemáticas e de Computação,Universidade de São Paulo. São Carlos, 2012 138p.

YANG , M.S. Survey of Fuzzy Clustering. Mathl. Comput. Modelling Vol. 18, No. 11, pp. 1-16, 1993

Xie, L. X.; Beni, G.. Validity measure for fuzzy clustering. IEEE Transactions on Pattern Analysis andMachine Intelligence, vol. 3, n. 8, p. 841-847, 1991.

Zadeh, L.A. (1965) Fuzzy sets, Inform. and Control 8, 338-353.