TEXTO PARA DISCUSSÃO Nº 1428

CLUSTERIZAÇÃO HIERÁRQUICA ESPACIAL COM ATRIBUTOS BINÁRIOS

Alexandre Xavier Ywata Carvalho
Pedro Henrique Melo Albuquerque
Gilberto Rezende de Almeida Junior
Rafael Dantas Guimarães
Camilo Rey Laureto

TEXTO PARA DISCUSSÃO Nº 1428

CLUSTERIZAÇÃO HIERÁRQUICA ESPACIAL COM ATRIBUTOS BINÁRIOS*

Alexandre Xavier Ywata Carvalho**
Pedro Henrique Melo Albuquerque***
Gilberto Rezende de Almeida Junior***
Rafael Dantas Guimarães***
Camilo Rey Laureto***

Brasília, outubro de 2009

* Os autores agradecem às sugestões de Bruno de Oliveira Cruz, Paulo Furtado de Castro e Bernardo Alves Furtado. Os erros remanescentes são de completa responsabilidade dos autores.
** Coordenador de Métodos Quantitativos da Diretoria de Estudos e Políticas Regionais, Urbanas e Ambientais (Dirur) do Ipea.
*** Pesquisadores-bolsistas do Programa de Pesquisa para o Desenvolvimento Nacional (PNPD) na Coordenação de Métodos Quantitativos da Dirur/Ipea.

Governo Federal

Secretaria de Assuntos Estratégicos da Presidência da República
Ministro Samuel Pinheiro Guimarães Neto

ipea **Instituto de Pesquisa Econômica Aplicada**

Fundação pública vinculada à Secretaria de Assuntos Estratégicos da Presidência da República, o Ipea fornece suporte técnico e institucional às ações governamentais – possibilitando a formulação de inúmeras políticas públicas e programas de desenvolvimento brasileiro – e disponibiliza, para a sociedade, pesquisas e estudos realizados por seus técnicos.

Presidente
Marcio Pochmann

Diretor de Desenvolvimento Institucional
Fernando Ferreira

Diretor de Estudos, Cooperação Técnica e Políticas Internacionais
Mário Lisboa Theodoro

Diretor de Estudos e Políticas do Estado, das Instituições e da Democracia (em implantação)
José Celso Pereira Cardoso Júnior

Diretor de Estudos e Políticas Macroeconômicas
João Sicsú

Diretora de Estudos e Políticas Regionais, Urbanas e Ambientais
Liana Maria da Frota Carleial

Diretor de Estudos e Políticas Setoriais, Inovação, Produção e Infraestrutura
Márcio Wohlers de Almeida

Diretor de Estudos e Políticas Sociais
Jorge Abrahão de Castro

Chefe de Gabinete
Persio Marco Antonio Davison

Assessor-chefe de Comunicação
Daniel Castro

URL: http://www.ipea.gov.br
Ouvidoria: http://www.ipea.gov.br/ouvidoria

ISSN 1415-4765

JEL J11, R11.

TEXTO PARA DISCUSSÃO

Publicação cujo objetivo é divulgar resultados de estudos direta ou indiretamente desenvolvidos pelo Ipea, os quais, por sua relevância, levam informações para profissionais especializados e estabelecem um espaço para sugestões.

SUMÁRIO

SINOPSE

Este trabalho discute uma metodologia para clusterização hierárquica espacial de polígonos contíguos e homogêneos de acordo com um conjunto de variáveis binárias. O algoritmo proposto é construído a partir de uma modificação do algoritmo aglomerativo de clusterização hierárquica tradicional, comumente utilizado na literatura de análise multivariada. De acordo com o método proposto neste estudo, a cada passo do processo sequencial de junção de *clusters*, impõe-se que somente conglomerados – grupos de polígonos originais, como municípios, estados ou setores censitários – vizinhos possam ser unidos para formar um novo *cluster* maior. Neste caso, foram definidos enquanto vizinhos polígonos que possuem um vértice em comum (vizinhança do tipo *queen*) ou uma aresta em comum (vizinhança do tipo *rook*). O texto apresenta aplicações da nova metodologia para clusterização dos municípios brasileiros, com base nas variações do número de empregos formais entre os anos de 1997 e 2007. Diversos métodos de clusterização são estudados, assim como diferentes tipos de distâncias entre vetores de variáveis binárias. Os métodos estudados foram: *centroid*, *single linkage*, *complete linkage*, *average linkage* e *average linkage weighted*, *Ward minimum variance* e método da mediana. As distâncias utilizadas foram: Jaccard, Tanimoto, *simple matching*, Russel e Rao, Dice, Kulczynski. Apresenta-se uma discussão sobre alguns métodos comumente aplicados para seleção do número de *clusters*. Finalmente, estudos de casos são apresentados para: *i)* comparar a formação dos algoritmos espaciais *versus* agrupamentos políticos existentes (microrregiões, mesorregiões e Unidades da Federação); e *ii)* identificar áreas no território brasileiro onde se verificou crescimento diversificado, em termos de atividades econômicas.

ABSTRACT

This paper studies a methodology for hierarchical spatial clustering of contiguous and homogeneous polygons, based on a set of binary variables. The proposed algorithm is built upon a modification of traditional agglomerative hierarchical clustering algorithm, commonly used in the multivariate analysis literature. According to the proposed method in this paper, at each step of the sequential process of collapsing clusters, only neighbor clusters (groups of original polygons, i.e. municipalities, census tracts, states) are allowed to be collapsed to form a bigger cluster. Two types of neighborhood are used: polygons with one edge in common (rook neighborhood) or polygons with only one point in common (queen neighborhood). In this paper, the methodology is employed to create clusters of Brazilian municipalities, based on the increase or decrease in the number of jobs between 1997 and 2007. Several clustering methods are investigated, as well as several types of vector distances for binary variables. The studied methods were: centroid method, single linkage, complete linkage, average linkage, average linkage weighted, Ward minimum variance e median method. The studied distances were: Jaccard, Tanimoto, simple matching, Russel e Rao, Dice, Kulczynski. A discussion on selection of the number of clusters is presented. Finally, case studies are presented in order to: (a) compare the intra-cluster variability of spatial hierarchical clusters versus the intra-cluster variability of existing political agglomerations (states, micro-regions and meso-regions); (b) identify areas or diversified economic growth.

1 INTRODUÇÃO

As últimas décadas têm testemunhado um grande avanço nas técnicas de tratamento de dados espaciais. Muitas destas técnicas têm sido importantes para auxiliar, por exemplo, na análise de heterogeneidades territoriais, que podem estar relacionadas a diversidades sociodemográficas, culturais, comportamentais e/ou a diversidades climáticas/geográficas. Tais diferenças espaciais podem surgir devido às grandes dimensões continentais (no caso de alguns países) e/ou a particularidades no histórico político/econômico de cada região. Esta heterogeneidade geográfica torna a execução de políticas públicas para o desenvolvimento regional uma tarefa complexa, exigindo técnicas, quantitativas e qualitativas, que forneçam uma desagregação espacial adequada, identificando áreas, dentro do território nacional, que possuam características semelhantes. Uma vez identificadas estas áreas, políticas específicas podem ser aplicadas.

Para políticas de desenvolvimento intraurbano, mesmo lidando com dimensões bem menores quando comparadas às dimensões do território nacional, o problema da heterogeidade espacial também está presente para todas as grandes cidades brasileiras. Políticas de combate à criminalidade, por exemplo, exigem estratégias diferentes para diferentes locais das grandes cidades, uma vez que as características dos crimes em cada local podem diferir sensivelmente (HIRSCHFIELD e BOWERS, 2001). Similarmente, políticas de melhorias habitacionais também devem levar em conta as particularidades sociodemográficas de cada área dentro de uma região metropolitana. Portanto, o problema de tratamento adequado da heterogeneidade espacial, do ponto de vista de políticas públicas, está presente também no caso de políticas intraurbanas.

Além das aplicações de técnicas de tratamento de heterogeneidade espacial de dados para estudos de desenvolvimento regional e estudos de desenvolvimento intraurbano, há uma gama de outras aplicações em outras áreas da ciência. No tratamento de dados de desmatamento, o pesquisador pode estar interessado em identificar áreas onde o desmastamento tenha sido mais intenso, nas últimas décadas. Estudos de epidemiologia exigem que o pesquiador possa detectar, com certa precisão, áreas de maior influência de uma determinadada enfermidade, de forma a desenvolver estratégias mais eficientes de combate. Por outro lado, estrategistas de campanhas de *marketing* podem estar interessados em dividir uma área metropolitana em subzonas homogêneas, de acordo com características socioeconômicas que influenciem os padrões de consumo. Com isso, é possível fazer um melhor direcionamento das ações publicitárias.

Uma das mais populares e eficientes técnicas de tratamento de dados de identificação de agregados homogêneos em um todo heterogêneo é a análise de *clusters* ou análise de agrupamentos. A técnica de tratamento de dados heterogêneos por meio de clusterização em grupos homogêneos é antiga, e está presente na maioria dos livros de estatística multivariada. A ideia de agrupar dados que compartilhem certas características vem desde a utilização de *clusters* unidimensionais, nos quais dados numa reta numérica são agrupados, até desenvolvimentos mais recentes na área de clusterização espacial. Berkhin (2002) traz uma linha do tempo explicando as diversas técnicas de clusterização. Para uma descrição geral dos algoritmos de clusterização, vide

Berry e Linoff (1997), Khattree e Naik (2000), Hastie, Tibshirani e Friedman (2001), Alpaydin (2004).

Especificamente para dados geográficos, a clusterização espacial é uma técnica poderosa, adaptável a diversos casos, e por isso vem ganhando espaço e desenvolvendo-se rapidamente. Grande parte destas técnicas de clusterização está baseada em modelos probabilísticos, para os quais procedimentos bayesianos ou de máxima verossimilhança são empregados. Li, Ramachandran, Movva Graves, Plale e Vijayakumar (2006), por exemplo, utilizaram técnicas de agrupamentos espaciais para aprimorar a previsão do tempo. Na área de saúde, Gangnon e Clayton (2003) estudaram casos de leucemia em Nova York, e, para aprimorar métodos de clusterização espacial, utilizam-se de simulações de Monte-Carlo e cadeias de Markov. Lawson e Denison (2002) apresentam uma coletânea de artigos sobre clusterização espacial aplicada a diversas áreas. Em geral, as técnicas baseadas em modelos probabilísticos tratam da identificação de áreas homogêneas com base na intensidade de ocorrências de eventos no espaço, ou com base em apenas uma variável de interesse (por exemplo, temperatura do ar e número de casos de leucemia por habitante).

Na linha de modelos probabilísticos para identificação de áreas com maior intensidade de ocorrência de eventos, está a análise de *hot spots*, muito popular em tratamento de dados de criminalidade (ECK *et al.*, 2005, HIRSCHFIELD e BOWERS, 2001). Neste caso, muito comumente os dados correspondem a coordenadas cartesianas (latitude e longitude) onde os crimes aconteceram, e são empregadas técnicas de estimação de densidades não paramétricas, em duas dimensões; os *hot spots* correspondem a áreas no espaço nos quais a densidade estimada tem valores mais altos. O objetivo da análise de *hot spots* não é dividir uma determinada cidade, por exemplo, em uma partição de subáreas homogêneas. O objetivo reside apenas em identificar locais onde a ocorrência de determinado evento é mais pronunciada.

Outra técnica comumente empregada para identificar áreas onde ocorrências de determinado evento são mais frequentes é a técnica *scan* (KULLDORFF, 1997; GLAZ e BALAKRISHNAN, 1999). Ao invés de informações sobre a localização geográfica (coordenadas cartesianas) das ocorrências, o método *scan* aplica-se a situações nas quais os dados estão disponíveis agregadamente por polígonos geográficos (municípios, setores censitários etc.). A técnica consiste basicamente em deslizar uma janela de tamanho fixo pelo espaço, em busca de uma região cuja densidade de pontos extrapola certo limite, comparando-se com alguma distribuição apropriada para o tipo de dados em questão (e.g. Bernoullli para dados binários ou Poisson para dados de contagem). Trata-se de uma técnica computacionalmente intensiva, e tem sido bastante utilizada para detecção geográfica de epidemias.

Neste texto, estuda-se uma metodologia para análise de dados espaciais conceitualmente diferente das técnicas de *hot spots* e das ténicas *scan*. Ao invés de tentar detectar áreas onde a ocorrência de determinado evento seja significantemente mais pronunciada, a técnica aqui abordada tem o objetivo de particionar a região de interesse (por exemplo, todo o território nacional) em sub-regiões, cada qual com características similares. O conjunto de informações analisado corresponde a: *i)* polígonos em um sistema georreferenciado (municípios, estados, setores censitários

etc.);e *ii)* variáveis características de cada polígono no sistema georeferenciado. Pode-se estar interessado em analisar a heterogeneidade espacial dos municípios brasileiros (polígonos), de acordo com renda *per capita*, longevidade, escolaridade média, condições dos domicílios (variáveis características). Denominou-se o método apresentado de *clusterização hierárquica espacial* ou *clusterização aglomerativa hierárquica espacial*, correspondente a uma modificação dos algoritmos de clusterização hierárquica tradicionais (HASTIE, TIBSHIRANI e FRIEDMAN, 2001; KHATTREE e NAIK, 2000; BERRY e LINOFF, 1997). Uma primeira motivação para a utilização deste algoritmo é a flexibilidade de incorporação de diferentes medidas de dissimilaridade, e de diferentes medidas de distância, entre vetores com variáveis contínuas, binárias, categóricas ou mistas.

Os métodos tradicionais de clusterização hierárquica consistem em identificar *clusters* homogêneos progressivamente, por meio da metamorfose (junção ou separação) de *clusters* anteriores na amostra. O critério para a formação progressiva de *clusters* é a distância entre eles. Diversas distâncias podem ser adotadas. Gower (1967) examina alguns métodos na análise de *clusters* e atenta para suas especificidades. A clusterização hierárquica pode ser feita de forma aglomerativa –iniciando com tantos *clusters* quantos objetos e então unindo-os em novos *clusters* – ou divisiva – iniciando com um *cluster* apenas e dividindo-o em novos *clusters*. A metamorfose de *clusters* é decidida por meio da proximidade entre objetos, fator de diferenciação entre os métodos de clusterização. A base do processo reside na construção da matriz de distâncias entre unidades observacionais ou conjuntos de unidades observacionais. No caso da clusterização hierárquica aglomerativa, os objetos próximos são unidos em *clusters* e a matriz de distâncias é atualizada. O processo interage até um número mínimo estabelecido de *clusters*.

De acordo com os algoritmos de clusterização hierárquica espacial estudados neste trabalho, os algoritmos de clusterização hierárquica tradicionais são modificados de forma a forçar a identificação de regiões geográficas, estritamente contíguas, com características socioeconômicas (ou segundo outro conjunto qualquer de variáveis) semelhantes. Entre as vantagens de se forçar a contiguidade, encontram-se:

1. O principal objetivo da análise de *clusters* é construir grupos homogêneos de áreas geográficas de acordo com um conjunto de variáveis (por exemplo, variáveis socioeconômicas). A hipótese implícita neste caso é que as variáveis utilizadas serão suficientes para descrever as características dos municípios ou setores censitários estudados. No entanto, pode acontecer que diversas outras variáveis, também importantes para a caracterização das unidades geográficas, não estejam incluídas na base, o que redundaria em alguma perda de informação na análise de clusterização. Por outro lado, pode-se esperar que as variáveis ausentes na base de dados apresentem uma forte correlação espacial, no sentido de que municípios vizinhos têm características semelhantes (ANSELIN, 1988; ANSELIN e FLORAX, 2000; PACE e BERRY, 1997). Neste caso, a utilização de algoritmos de clusterização onde a contiguidade é imposta pode reduzir a perda de informação devido à ausência de algumas variáveis na base de dados.

2. Especificamente para trabalhos nas áreas de desenvolvimento regional e intraurbano, por exemplo, o principal objetivo é identificar regiões homogêneas, no país ou dentro de uma área urbana, onde políticas de desenvolvimento diferenciadas possam ser implementadas. Dessa forma, a contiguidade é fundamental, pois a intenção é a formulação de políticas públicas focadas para áreas geográficas que apresentem algum grau de vizinhança.

Este texto corresponde a uma continuação do estudo de Carvalho *et al.* (2009), que investigam o método de clusterização espacial hierárquica. No caso, os autores objetivam o tratamento de bases de dados com atributos contínuos, para os quais são utilizadas diferentes distâncias topológicas. As distâncias utilizadas no estudo citado são: distância Euclidiana (norma L_2), norma L_1 – distância de Manhattan, norma L_p (caso mais geral), distância de Mahalanobis e distância euclidiana corrigida pela variância (*variance corrected*). O presente trabalho, por sua vez, considera exclusivamente variáveis binárias (com valores zero ou um). Pretende-se assim estudar o efeito de diferentes tipos de distâncias para atributos binários, combinados a diferentes medidas de dissimilaridade. Em etapas posteriores do projeto de pesquisa, pretende-se criar uma metodologia mais geral, para tratar bases com variáveis de diferentes tipos (contínuas, binárias, categóricas etc.), seguindo os princípios propostos em Hiu *et al.* (2001).

Outra diferença em relação a Carvalho *et al.* (2009) refere-se a artifícios utilizados neste texto para evitar que um único *cluster* resulte em um número muito grande de unidades geográficas. Conforme observado no primeiro estudo, para alguns dos métodos utilizados uns poucos *clusters* continham quase todos os polígonos no sistema georreferenciado (municípios brasileiros). Do ponto de vista de análise geográfica, isto pode não ser muito útil, tendo em vista que a existência de diferenças muito grandes nos tamanhos dos *clusters* pode resultar em *clusters* pouco úteis do ponto de vista de estratégia de políticas públicas, por exemplo. Para contornar este problema e garantir que todos os métodos resultem em *clusters* com tamanhos não muito discrepantes, utilizaram-se, no presente texto, mecanismos de forma que a distância dos *clusters*, com um número de unidades acima de um valor de corte aos demais fosse deliberadamente inflacionada para evitar que estes *clusters*, já considerados grandes, se unissem a outros polígonos.

Os três desafios da clusterização são estabelecer a medida de similaridade ou de dissimilaridade empregada, estabelecer a distância entre vetores de dados e definir o número final de *clusters*. As medidas de dissimilaridade empregadas neste trabalho são: *average linkage*, *centroid*, *single linkage*, *complete linkage (unweighted), complete linkage (weighted)*, *ward minimum variance* e método da mediana. Além das diferentes medidas de dissimilaridade, empregaram-se também diferentes distâncias entre vetores com variáveis binárias: Jaccard, Tanimoto, *simple matching*, Russel e Rao, Dice, Kulczynski. Por fim, será feita a discussão a respeito da definição do número de *clusters*, utilizando os critérios *CCC*, pseudo-*F*, pseudo-t^2, R^2 e R^2 semiparcial. Apesar de estes critérios terem sido construídos especificamente para variáveis contínuas, a ideia é checar como eles se comportam no caso dos algoritmos para variáveis binárias propostos neste estudo.

Algoritmos alternativos para construção de *clusters* espaciais – ou seja, com *clusters* compostos por unidades contíguas – estão descritos, por exemplo, em Maravalle e Simeone (1995) e Maravalle, Simeone e Naldini (1997). Estes autores propõem algoritmos baseados na transformação de um mapa em um grafo, e na posterior redução do grafo a uma árvore geradora mínima. Aplicações destes algoritmos de clusterização a partir de grafos para o Brasil estão apresentadas em Assunção, Lage e Reis (2002), e Chein, Lemos e Assunção (2005). Apesar de os algoritmos apresentados em Maravalle e Simeone (1995) e Maravalle, Simeone e Naldini (1997) terem os mesmos objetivos de análise que os algoritmos de clusterização espacial hierárquica, tratados neste texto, os algoritmos hierárquicos aqui apresentados possuem as seguintes diferenças em relação aos *clusters* espaciais via árvore geradora mínima:

a) Os algoritmos via árvore geradora mínima baseiam-se em uma sequência divisiva de passos; inicia-se com um único grafo, e reparte-se este grafo sequencialmente, obtendo-se novos subgrafos. Por sua vez, o algoritmo de clusterização espacial estudado neste texto baseia-se em uma sequência aglomerativa de passos.

b) Os algoritmos de clusterização hierárquica espacial permitem a incorporação de diferentes medidas de dissimilaridade (Ward, *simple linkage*, *complete linkage*, *average linkage*, *average linkage weighted*, mediana, *centroid*) entre grupos homogêneos e diferentes distâncias para diversos tipos de variáveis (contínuas, binárias, categóricas).

c) Os algoritmos neste trabalho permitem a utilização de critérios tradicionais de escolha do número (estatísticas *CCC*, *pseudo-F* e pseudo-t^2, R^2 e R^2 semiparcial) de *clusters* em algoritmos aglomerativos sequenciais.

d) Os algoritmos de clusterização espacial constituem-se em uma abordagem mais intuitiva, na qual os passos dos algoritmos ficam nítidos tanto para os usuários da nova metodologia quanto para os leitores.

Uma discussão extensa sobre métodos de classificação com restrições pode ser encontrada em Batagelj e Ferligoj (1998), Gordon (1996), Duque, Ramos e Surinach (2007). Um algoritmo de clusterização considerando-se contiguidade espacial é fornecido em Luo (2001); neste artigo, o autor sugere a utilização do algoritmo K-*means*, no qual a função objetivo contém um termo para penalizar a ausência de contiguidade entre unidades clusterizadas. A ideia de restringir as aglomerações em algoritmos de clusterização hierárquica foi implementada em Wipperman (1999); neste artigo, o autor considera apenas duas medidas de dissimilaridade: *single linkage* e *complete linkage*, utilizando a distância euclidiana entre os vetores de atributos, e aplica o algoritmo para a construção de territórios homogêneos para regras de seguros de veículos na província de British Columbia, Canadá. A aplicação apresentada em Wipperman (1999) considerou apenas 11 áreas geográficas, e uma análise detalhada de cada passo aglomerativo é apresentada. O presente texto complementa a literatura com os seguintes itens:

1. Consideram-se exclusivamente medidas de distâncias para vetores de variáveis binárias, complementando as análises específicas para variáveis contínuas, apresentadas em Carvalho *et al.* (2009).
2. Uma comparação entre as diferentes definições de vizinhança entre polígonos é discutida.
3. Diversos tipos de medidas de dissimilaridade e diversos tipos de distâncias entre vetores binários são considerados – a combinação entre as medidas de dissimilaridade e as distâncias entre vetores possibilita a criação de uma variadade de algoritmos diferentes.
4. Apresenta-se uma utilização dos algoritmos estudados para dados reais, de crescimento ou decréscimo no número de empregos formais, para diferentes setores de atividade econômica, nos municípios brasileiros, entre 1997 e 2007. Este exercício permite a identificação de áreas no território brasileiro onde se observa crescimento da economia, com diversificação das atividades.
5. Introduz-se um artifício no algoritmo de construção dos *clusters* hierárquicos, de forma a penalizar a formação de *clusters* com muitas unidades geográficas – o que poderia ser indesejável no caso de construção de agrupamentos espaciais.

O trabalho está dividido em quatro seções, incluindo esta introdução. A segunda parte aborda a metodologia utilizada para formação dos grupos homogêneos de municípios (*clusters*), discussão sobre as medidas de dissimilaridade, cálculo das diferentes distâncias, e discussão sobre os critérios de seleção do número de *clusters*. A terceira seção apresenta um exercício de comparação entre as diversas medidas de dissimilaridade e as diversas distâncias utilizadas, com base em um estudo de caso, utilizando-se variações positivas (um) ou negativas (zero) no número de postos de trabalho, por divisões CNAE 95 (Classificação Nacional de Atividades Econômicas), nos municípios brasileiros, entre os anos de 2007 e 2008. A terceira seção apresenta também os resultados de estudos de casos para: *i)* comparar a formação dos algoritmos espaciais *versus* a agrupamentos políticos existentes (microrregiões, mesorregiões e Unidades da Federação); e *ii)* identificar áreas no território brasileiro onde houve crescimento diversificado, em termos de atividades econômicas. A quarta seção é reservada para as conclusões do trabalho.

2 METODOLOGIA

Nesta seção, descreve-se o algoritmo para formação dos agrupamentos homogêneos de municípios, no caso de variáveis binárias. Conforme será abordado em mais detalhes, o algoritmo utilizado neste trabalho corresponde a uma modificação dos algoritmos de clusterização hierárquica comumente expostos na literatura de análise multivariada.

2.1 ALGORITMO AGLOMERATIVO PARA FORMAÇÃO DE GRUPOS ESPACIAIS HOMOGÊNEOS

Nos algoritmos tradicionais de clusterização (hierárquica ou não), quando são agrupadas unidades geográficas do tipo municípios ou setores censitários, não necessariamente os grupos homogêneos são formados por municípios ou setores

censitários estritamente vizinhos. Pode acontecer que, em um mesmo *cluster*, haja municípios geograficamente separados. A formação de agrupamentos homogêneos de municípios, com componentes não necessariamente contíguos, pode não ser um problema em muitas das aplicações. De fato, pode acontecer de o analista ou pesquisador estar interessado justamente em identificar se existem regiões (setores censitários, áreas de ponderação) na periferia de São Paulo, por exemplo, que são semelhantes, em termos de atributos socioeconômicos, a regiões no centro da cidade.

A seguir, apresenta-se uma descrição sucinta dos algoritmos de clusterização hierárquica tradicionais. Em seguida, discutem-se as modificações no método de clusterização tradicional, de forma a incorporar a restrição de unidades geográficas contíguas (a exemplo de municípios, setores censitários, Unidades da Federação, áreas de ponderação).

2.2 ALGORITMOS DE CLUSTERIZAÇÃO HIERÁRQUICA

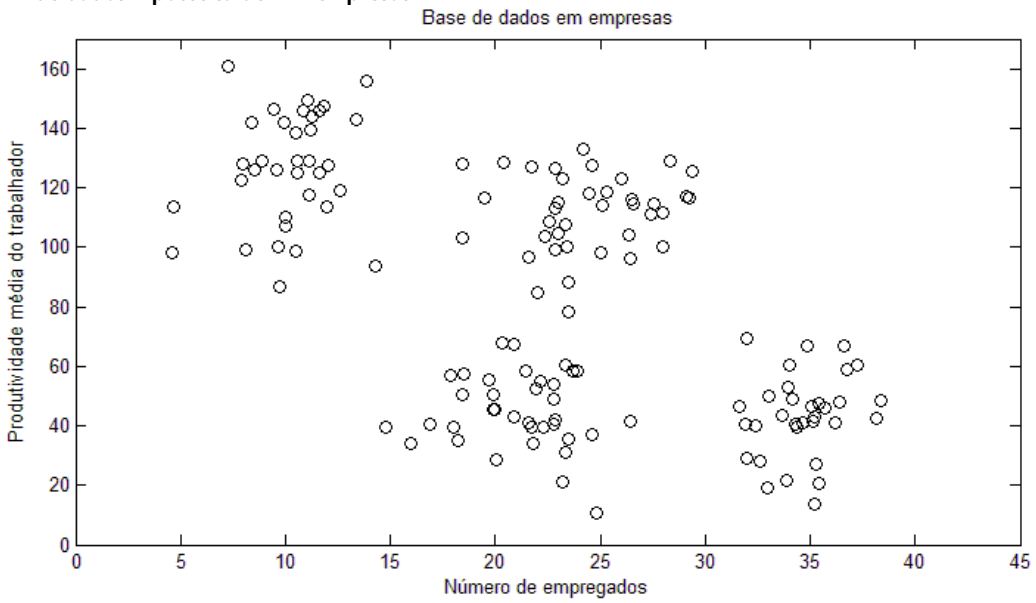
Para exemplificar a ideia geral dos algoritmos de clusterização, considere-se a figura 2.1, que contém 141 observações, cada qual correspondente a uma empresa específica (base de dados hipotética, para fins de ilustração). O eixo horizontal do gráfico indica o número de empregados em cada uma das empresas, enquanto o eixo vertical indica a produtividade média dos trabalhadores.

Por meio de uma análise visual simples, as 141 empresas podem ser divididas em quatro grupos homogêneos em relação às duas variáveis: número de empregados e produtividade marginal dos trabalhadores. Estes grupos estão melhor representados na figura 2.2. Observe-se que o grupo 1, em vermelho,[1] pode ser interpretado como o grupo de empresas com baixo número de empregados e alta produtividade. O grupo 2, em azul, pode ser considerado o grupo de baixo número de empregados e baixa produtividade. O grupo 3, em preto, corresponde às empresas de médio porte, e com alta produtividade. Finalmente, o grupo 4, em verde, corresponde às empresas com muitos empregados e baixa produtividade.

Nesse exemplo, para a amostra de 141 empresas hipotéticas, a identificação dos grupos homogêneos é trivial, podendo ser executada por um simples procedimento gráfico. No entanto, na grande maioria dos problemas práticos, procedimentos gráficos têm aplicabilidade limitada. Os principais complicadores são: *i)* as bases de dados podem conter um número muito grande de observações (não raramente na casa dos milhões); *ii)* o número de variáveis de caracterização das observações é bem superior a três, impossibilitando a confecção de gráficos dos pontos na amostra, mesmo em três dimensões; e *iii)* a determinação do número de *clusters* (grupos homogêneos) não é tão imediata.

1. Para visualização em cores, acessar a seção *O trabalho do Ipea*, subseção *Publicações*, no site: <http://www.ipea.gov.br>.

FIGURA 2.1
Informações sobre número de empregados e produtividade média do trabalhador para uma base de dados hipotética de 141 empresas

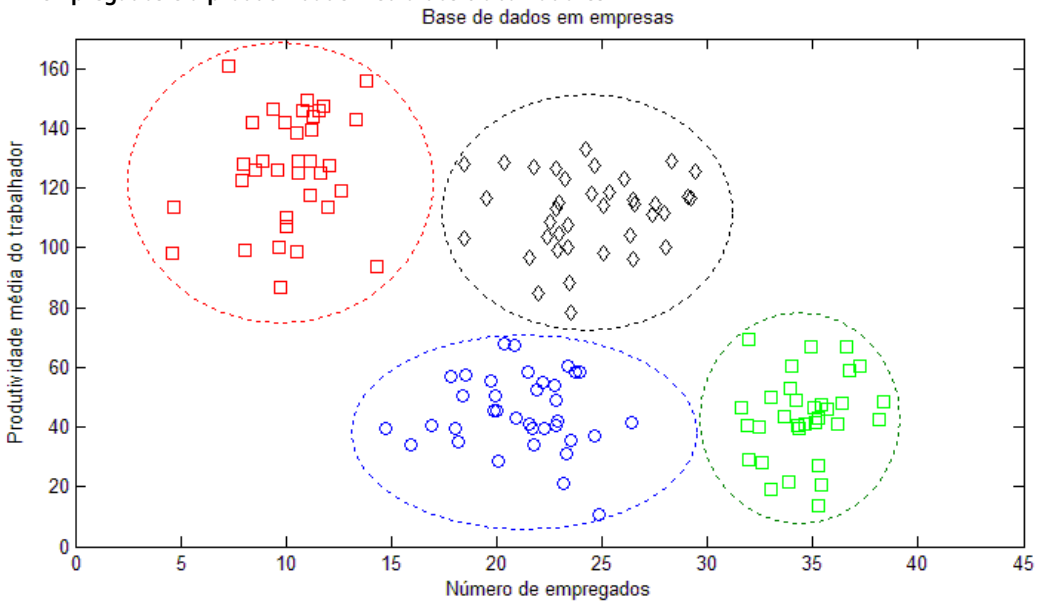


Elaboração dos autores.

Dada a grande importância do problema de clusterização de observações, pesquisadores em estatística, matemática aplicada e ciência da computação têm se dedicado à construção de algoritmos computacionais que possam realizar automaticamente o que foi feito no problema acima de forma visual. Hastie, Tibishirani e Friedman (2001) apresentam uma descrição geral destes algoritmos. Os algoritmos de clusterização podem ser divididos em três grandes categorias: *i)* algoritmos combinatórios (*combinatorial algorithms*); *ii)* misturas de modelos (*mixture models*); e *iii)* busca por modas (*mode seeking*). As últimas duas categorias baseiam-se em alguma forma de modelos probabilísticos para o processo gerador de dados. Por seu turno, os algoritmos combinatórios podem ser vistos basicamente enquanto regras heurísticas de busca dos melhores agrupamentos de observações. De forma geral, não existe um algoritmo superior aos demais em todas as situações. Qual deles melhor se aplica dependerá do processo gerador de dados, bem como da experiência do analista ou pesquisador, e da disponibilidade de *softwares* específicos.

O algoritmo empregado neste trabalho pode ser classificado como um algoritmo combinatório, e tem uma estrutura de formação de *clusters* do tipo hierárquica aglomerativa. Para uma descrição mais detalhada deste tipo de metodologia, vide Khattree e Naik (2000). De maneira geral, o algoritmo tradicional tem os seguintes passos:

FIGURA 2.2
Identificação visual de quatro grupos homogêneos de empresas, de acordo com o número de empregados e a produtividade média dos trabalhadores.



Elaboração dos autores.

1. Considere-se uma base inicial de N *clusters*. Em geral, estes agrupamentos correspondem simplesmente às unidades a serem agrupadas em subgrupos homogêneos (por exemplo, empresas, clientes, municípios etc.). Portanto, em geral, cada um destes N *clusters* contém inicialmente apenas uma unidade. A cada unidade i, está associado um vetor de m características $x_i = [x_{i,1}\ x_{i,2}\ \ldots\ x_{i,m}]$. Estas características podem ser características socioeconômicas, por exemplo. Neste trabalho, o vetor de características é composto exclusivamente de variáveis binárias, assumindo valor zero ou um.

2. Calcula-se a distância entre todos os pares formados por elementos entre esses N *clusters* iniciais. Distância, neste caso, pode ser qualquer medida de dissimilaridade entre o conjunto de atributos binários $x_i = [x_{i,1}\ x_{i,2}\ \ldots\ x_{i,m}]$. Para uma discussão sobre as diversas medidas de dissimilaridade, vide Khattree e Naik (2000) e Berry e Linoff (1997). Entre as diversas medidas de dissimilaridade possíveis, cite-se a medida de McQuitty, que pode ser escrita como:

$$D_{K,L} = \frac{1}{N_K N_L} \sum_{i \in C_K} \sum_{i \in C_L} d(x_i, x_j)$$

onde $D_{K,L}$ é a medida de dissimilaridade entre o *cluster* L e o *cluster* K, C_K e C_L são os conjuntos de elementos dentro dos *clusters* L e K, $d(x_i,x_j)$ é uma distância entre os vetores, de variáveis binárias, x_i e x_j, e N_L e N_K são as quantidades de unidades dentro dos *clusters* L e K (lembrando que cada *cluster* pode conter mais de uma unidade). A seção 2.4 apresenta um conjunto de outras medidas de dissimilaridade (além de uma discussão mais detalhada sobre a medida de dissimilaridade de McQuitty), bem assim um conjunto de distâncias entre vetores de variáveis binárias, utilizadas neste texto.

3. Sejam *I* e *J* os dois *clusters* apresentando a menor distância, ou dissimilaridade, entre eles. Agrupa-se então o par *I* e *J* em um único novo *cluster*. O número de *clusters* agora passa a ser *N-1*.

4. Para os *N-1* novos *clusters*, depois da junção descrita no passo 3, calculam-se as distâncias entre todos os pares. Para o par com a menor distância, agrupam-se os elementos em um único novo *cluster*, de forma que o número de *clusters* existentes passe a ser *N-2*.

5. Repetem-se os passos 2 a 4 até se obter um único *cluster*, que deverá conter todos os *N* clusters iniciais.

Ao final do processo, o analista terá em mãos uma árvore descrevendo a sequência de agrupamentos em cada passo do algoritmo. Para um número inicial de *N* unidades observacionais na base de dados, ao todo ocorrem *N-1* junções. Diversos *softwares* estatísticos apresentam recursos gráficos que permitem ao usuário apresentar a árvore construída.

Algoritmos hierárquicos em geral, conforme apresentado acima, são muito demandantes computacionalmente. Na primeira iteração do processo, o número de pares de observações possíveis é igual $N \times (N-1) / 2$. Na segunda iteração, o número de pares passa a ser $(N-1) \times (N-2) / 2$, o que ainda pode ser um número elevado. Para bases de dados com muitas observações, a implementação de algoritmos hierárquicos, de acordo com os passos acima, torna-se impossível. Nestas situações, diversas alternativas existem – por exemplo, o sorteio de uma subamostra das *N* observações para posterior comparação. No entanto, para situações envolvendo unidades geográficas, como municípios ou setores censitários, o número de unidades *N* não é tão grande, e o algoritmo original pode ser empregado com recursos computacionais comumente disponíveis. Além disso, conforme o passo 2 do algoritmo de clusterização hierárquica espacial, descrito a seguir, em cada iteração do algoritmo o número de pares de observações comparadas não mais será $N \times (N-1) / 2$, dado que as comparações serão feitas apenas entre unidades geográficas vizinhas. Isso reduz enormemente o tempo de processamento.

O passo final é então selecionar o número de *clusters* ou de grupos homogêneos. No exemplo acima, quatro parece ser um número graficamente adequado. No entanto, na maioria das situações práticas, a escolha do número de *clusters* não é tão simples. Diversas medidas estatísticas para seleção do número de agrupamentos foram propostas, sem haver necessariamente um consenso sobre qual medida utilizar. Algumas dessas estatísticas são a *CCC*, a pseudo-*F* e a pseudo-t^2 (KHATTREE e NAIK, 2000). De maneira geral, estas medidas estão associadas a um indicador de dissimilaridade agregada entre todos os *clusters* construídos. Por meio de um gráfico destas medidas *versus* o número de *clusters* selecionado, é possível identificar aumentos expressivos (picos) no grau de dissimilaridade para algum número específico de *clusters*. Estes picos no grau de dissimilaridade agregada sugerem então pontos de parada no algoritmo de agregação sequencial apresentados nos passos 1 a 5 acima, indicando portanto quantos *clusters* utilizar. A seção 2.5 apresenta uma discussão sobre alguns dos critérios de seleção do número de *clusters* comumente empregados. Por outro lado, para estudos com bases de dados de informações socioeconômicas, é interessante ter uma interpretação plausível para todos os *clusters* formados. Isso

permite combinar algoritmos computacionais robustos com a informação do analista, que sempre deve ser levada em conta.

2.3 ALGORITMOS DE CLUSTERIZAÇÃO HIERÁRQUICA ESPACIAL

Os algoritmos de clusterização mais comuns foram desenvolvidos visando a aplicações em diferentes áreas, nas quais as unidades observacionais podem ser de diversas naturezas. Em estudos de *marketing*, as unidades agrupadas geralmente são clientes ou compradores. Em estudos genéticos, as unidades clusterizadas podem ser sequências de DNA, por exemplo. Neste trabalho, as unidades a serem agrupadas são municípios ou setores censitários, por exemplo, e os *clusters* correspondem a regiões de municípios ou setores homogêneos, para as quais políticas de desenvolvimento regional ou urbano específicas possam ser propostas. Espera-se que os *clusters* formados sejam compostos de unidades geográficas homogêneas e vizinhas. Por outro lado, a aplicação direta de qualquer um dos algoritmos comumente encontrados na literatura, e disponíveis em pacotes estatísticos, muito provavelmente fornecerá grupos homogêneos formados por unidades geográficas que não apresentem contiguidade entre elas. Pode acontecer, por exemplo, que um mesmo grupamento contenha um município localizado ao sul do estado e outro município localizado no extremo norte. Nesta seção, discutem-se algumas modificações impostas no algoritmo de clusterização hierárquica apresentado na seção anterior, de forma a incorporar explicitamente a restrição de contiguidade entre as unidades geográficas que compõem um mesmo *cluster*.

Os passos a seguir descrevem a modificação do algoritmo hierárquico, de forma a satisfazer a restrição de vizinhança entre as unidades de cada agrupamento homogêneo. Para facilitar a apresentação, os passos descritos referem-se à clusterização de municípios; porém, a discussão é imediatamente aplicável para qualquer outro tipo de unidade geográfica.

1. Seja C uma base inicial de N unidades geográficas, que já podem corresponder a agrupamentos iniciais de subunidades (no estudo empírico apresentado neste texto, estas subunidades correspondem a municípios). Inicialmente, cada uma destas N observações consiste em um *cluster* isoladamente, e tem um conjunto de atributos (variáveis) $[x_{i,1}\ x_{i,2}\ \ldots\ x_{i,m}]$. Para cada uma destas N unidades, é preciso determinar a lista de unidades vizinhas, de acordo com algum critério espacial. Neste projeto, foram investigadas duas definições de vizinhança. No primeiro caso, foram considerados municípios vizinhos aqueles que possuem pelo menos um lado em comum (considerando-se um sistema de dados georreferenciados) – este tipo de vizinhança é conhecida na literatura de estatística espacial como vizinhança do tipo *rook*. No segundo caso, foram considerados municípios vizinhos aqueles que possuem pelos menos um vértice em comum – este tipo de vizinhança é denominado vizinhança do tipo *queen*. Obviamente, a vizinhança do tipo *queen* é menos restritiva do que a vizinhança do tipo *rook*.
2. Calcula-se a medida de dissimilaridade entre todos os pares formados por elementos estritamente vizinhos na lista de N unidades. A seção 2.4 apresenta uma descrição das medidas de dissimilaridade e das distâncias entre vetores

utilizadas no estudo empírico. O número de pares testados não é mais $N\times(N\text{-}1)/2$, como no algoritmo hierárquico tradicional, já que nem todos os pares são formados por unidades geográficas vizinhas. Portanto, a restrição de contiguidade possibilita a construção de algoritmos com tempo de processamento bem menor.

3. Sejam *I* e *J* as duas unidades geográficas vizinhas apresentando a menor distância, ou dissimilaridade, entre elas. Agrupa-se o par *I* e *J* em um único *cluster*. O número de *clusters* agora passa a ser *N*-1.

4. Na definição do novo *cluster*, formado pelas unidades *I* e *J*, serão combinadas não somente as listas de atributos binários $[x_{i,1}\ x_{i,2}\ \ldots\ x_{i,m}]$, mas também as listas de vizinhos. Portanto, será composta uma nova lista de municípios vizinhos a partir da união da lista de vizinhos do município *I* com a lista de vizinhos do município *J*.

5. Para os *N*-1 novos *clusters*, depois da junção descrita nos itens 3 e 4, calculam-se as medidas de dissimilaridade entre todos os pares de *clusters* vizinhos. Neste caso, dois *clusters* A e B de municípios são considerados vizinhos quando houver pelo menos um município em A que é vizinho de um município em B. Para o par de *clusters* com a menor distância, agrupam-se os elementos em um único novo *cluster*, de forma que o número de *clusters* existentes passe a ser *N*-2. Ressalta-se que a distância entre os *clusters* A e B corresponde unicamente à dissimilaridade entre os atributos binários $[x_{i,1}\ x_{i,2}\ \ldots\ x_{i,m}]$. Em geral, estas variáveis correspondem a características socioeconômicas – caso do estudo empírico neste trabalho – e não contêm necessariamente informações sobre localização geográfica. A similaridade geográfica já está explicitamente modelada quando são agrupados somente *clusters* vizinhos.

6. Repetem-se os passos 2 a 5 até se obter um único *cluster*, que deverá conter todas as *N* unidades geográficas originais.

Da mesma forma que no caso da clusterização hierárquica tradicional, ao final do processo tem-se uma árvore caracterizando os agrupamentos decorridos em cada passo do algoritmo. Novamente, o analista pode recorrer a alguns dos indicadores tradicionais (por exemplo, *CCC*, pseudo-*F* e pseudo-t^2) para a escolha do número de agrupamentos mais apropriado. No entanto, devido ao fato de o algoritmo utilizado neste estudo ser completamente original e utilizar modificações substanciais nos algoritmos de clusterização hierárquica tradicionais, as propriedades destes indicadores estatísticos não necessariamente são as mesmas propriedades para a clusterização tradicional (não espacial), havendo a necessidade de estudos posteriores do comportamento de tais indicadores. Além disso, estes critérios foram criados para a avaliação do número de *clusters* a partir de variáveis contínuas. Pesquisas futuras são necessárias para construção de critérios de seleção de variáveis no caso de variáveis binárias.

A utilização direta de critérios estatísticos não necessariamente implicará em um número de *clusters* que faça sentido de acordo com os objetivos de cada projeto. Pode-se optar por selecionar o número de agrupamentos cuja interpretação econômica faça mais

sentido de acordo com os objetivos do trabalho. A escolha do número de agrupamentos homogêneos via critérios subjetivos foi utilizada, por exemplo, em Chein, Lemos e Assunção (2005), que selecionaram 100 *clusters* para todo o território brasileiro.

2.4 MEDIDAS DE DISSIMILARIDADE ENTRE *CLUSTERS*

Nesta seção, apresentam-se algumas das medidas de dissimilaridade comumente encontradas na literatura de clusterização hierárquica. Estas medidas de dissimilaridade definem o método de clusterização hierárquica empregado. A lista de medidas abaixo não é exaustiva, podendo o leitor recorrer às referências apresentadas neste texto para conhecer outras medidas. Além das medidas de dissimilaridade apresentadas a seguir, que serão investigadas no estudo de caso apresentado mais adiante, esta seção apresenta uma lista de distâncias entre vetores com variáveis binárias. Estas distâncias podem ser combinadas com as medidas de dissimilaridade, para gerar uma grande variedade de métodos possíveis. As várias combinações são estudadas na seção 3.2.

2.4.1 Medidas de Dissimilaridade

As medidas de dissimilaridade aqui estudados estão listadas a seguir. Além das expressões para cada medida, faz-se uma breve discussão sobre o comportamento observado em aplicações para clusterização hierárquica tradicional.

Average linkage (unweighted)

O método *average linkage* – também conhecido por método de McQuitty – define a distância média entre pares de objetos enquanto sendo a relevante para elaboração da matriz de distâncias. É um método que tende a juntar *clusters* com baixa variância, e é ligeiramente viesado a produzir *clusters* com igual variância. A medida de dissimilaridade entre os *clusters* K e L é definida por

$$D_{K,L} = \frac{1}{N_K N_L} \sum_{i \in C_K} \sum_{i \in C_L} d(x_i, x_j)$$

Outra maneira de implementar o algoritmo com essa medida de dissimilaridade é através da atualização da matriz de dissimilaridade entre *clusters*. Toda vez que um novo *cluster* C_M é criado, a partir da junção de *clusters* C_L e C_K existentes no passo anterior, a matriz de dissimilarides é atualizada, para considerar as distâncias ao novo *cluster*. Esta atualização pode ser feita diretamente a partir das distâncias existentes na matriz no passo anterior, utilizando-se fórmulas combinatórias (*combinatorial formulas*). Considere-se então um *cluster* qualquer C_J. A dissimilaridade entre C_J e o novo *cluster* C_M pode ser obtida a partir das dissimilaridades anteriores, utilizando-se a expressão

$$D_{J,M} = \frac{1}{2} D_{J,K} + \frac{1}{2} D_{J,L}$$

O método *average linkage* considera uma média de todos os membros dos *clusters* cuja distância está sendo calculada. Consequentemente, este método é menos influenciado por valores extremos – o que ocorre com os métodos de *single linkage* e *complete linkage*.

Single linkage

O método *single linkage* (ou do vizinho mais próximo) baseia-se na distância entre os pontos de cada *cluster* mais próximos entre si. Possui muitas propriedades teóricas desejáveis, mas manifesta mau desempenho em experimentos de Monte-Carlo. A medida de dissimilaridade é definida por:

$$D_{K,L} = \min_{i \in C_K, j \in C_L} d(x_i, x_j)$$

A dissimilaridade entre um *cluster* qualquer C_J e o novo cluster C_M pode ser obtida utilizando-se a fórmula combinatória

$$D_{J,M} = \frac{1}{2} D_{J,K} + \frac{1}{2} D_{J,L} - \frac{1}{2} \left| D_{J,K} - D_{J,L} \right|$$

Tendo em vista que não impõe restrições na forma dos *clusters*, esse método sacrifica a possibilidade de obter *clusters* compactos com a vantagem de permitir a obtenção de *clusters* irregulares ou alongados. O *single Linkage* também tende a cortar as caudas das distribuições antes de separar os *clusters* principais. A evidente tendência de encadeamento do *single linkage* pode ser aliviada.

Complete linkage method

O método *complete linkage* baseia-se na distância entre os pontos de cada *cluster* mais distantes entre si. É fortemente viesado no sentido de produzir *clusters* compactos com diâmetros semelhantes, e pode ser severamente distorcido por *outliers* moderados. É um método que assegura que todos os itens de um *cluster* estão a uma distância mínima um do outro.

$$D_{K,L} = \max_{i \in C_K, j \in C_L} d(x_i, x_j)$$

A dissimilaridade entre um *cluster* qualquer C_J e o novo *cluster* C_M pode ser obtida utilizando-se a fórmula combinatória

$$D_{J,M} = \frac{1}{2} D_{J,K} + \frac{1}{2} D_{J,L} + \frac{1}{2} \left| D_{J,K} - D_{J,L} \right|$$

Este método é severamente influenciado por valores discrepantes.

Método de mínima variância de Ward

O método de mínima variância de Ward é viesado na direção de gerar *clusters* de igual tamanho. O método parte da soma dos quadrados dos erros (*SQE*) de cada *cluster* (soma dos quadrados dos desvios para o centroide do *cluster*). Somam-se os *SQE*s de todos os *G clusters*, gerando o *SQET*. O método consiste em analisar todos os possíveis pares de *cluster* unidos, detectando qual união produz o menor aumento de *SQE*. Neste método, a distância entre dois *clusters* é a soma de quadrados ANOVA entre os dois *clusters*, para todas as variáveis. A cada geração, minimiza-se a soma de quadrados intracluster obtenível pela união de dois *clusters*. Frequentemente, aconselha-se utilizar a razão *SQE*/*SQET* no lugar do *SQE* absoluto. A fórmula combinatória é dada por:

$$D_{J,M} = \frac{n_J + n_K}{n_J + n_L + n_K} D_{J,K} + \frac{n_J + n_L}{n_J + n_L + n_K} D_{J,L} - \frac{n_J}{n_J + n_L + n_K} D_{K,L}$$

Trata-se de um método que busca maximizar a verossimilhança em cada nível de hierarquia sob as hipóteses de mistura de normais multivariadas, matrizes esféricas de covariância iguais e probabilidades amostrais iguais. Tende a unir *clusters* com número pequeno de observações e é fortemente viesado na direção de produzir *clusters* com mesmo formato e número de observações. É também muito sensível a *outliers*.

Centroid method

Desenvolvido por Sokal e Michener em 1958, o método *centroid linkage* considera a distância entre *clusters* o quadrado da distância euclidiana entre os centroides dos *clusters*. A fórmula combinatória para este método é dada por:

$$D_{J,M} = \frac{n_K}{n_L + n_K} D_{J,K} + \frac{n_L}{n_L + n_K} D_{J,L} - \frac{n_L n_K}{(n_L + n_K)^2} D_{K,L}$$

Por tratar-se de uma comparação de médias, *outliers* exercem pouca influência. Em outros aspectos, pode perder em eficiência para os métodos de *Average Linkage* e Ward. O maior dos dois *clusters* unidos tende a dominar o novo *cluster*.

Average linkage weigthed

O método *average linkage* ponderado diferencia-se do método *average linkage* original devido aos pesos diferenciados inseridos na fórmula combinatória. A nova expressão combinatória passa a ser:

$$D_{J,M} = \frac{n_K}{n_L + n_K} D_{J,K} + \frac{n_L}{n_L + n_K} D_{J,L}$$

Na expressão, n_K e n_L são os números de observações nos *clusters* K e L respectivamente.

Método da mediana

O método da mediana tem expressão para atualização das distâncias da matriz de dissimilaridade dada por:

$$D_{J,M} = \frac{1}{2} D_{J,K} + \frac{1}{2} D_{J,L} - \frac{1}{4} D_{K,L}$$

Esse método foi desenvolvido por Gower (1967).

2.4.2 Tipos de Distâncias

São apresentados, a seguir, os tipos de distâncias para vetores de variáveis binárias que serão utilizados neste texto. Em matemática, uma métrica ou função distância é uma função que define uma distância entre elementos de um determinado conjunto. Um conjunto com uma métrica é denominado espaço métrico. Sejam x e y vetores contendo as variáveis de caracterização para dois polígonos quaisquer. No caso de dados municipais (espaço métrico), por exemplo, x e y podem corresponder a vetores contendo as variáveis binárias referentes a crescimento ou decréscimo do número de empregos formais. Será utilizada a notação x_i para especificar o i-ésimo elemento (escalar) do vetor x.

A tabela 2.1 apresenta uma lista das distâncias entre variáveis binárias utilizadas neste trabalho. A segunda coluna apresenta as expressões, com base nas grandezas definidas pelas equações a seguir. O valor v corresponde ao número de variáveis em cada vetor.

$$a_1 = \sum_{k=1}^{v} I(x_k = y_k = 1)$$

$$a_2 = \sum_{k=1}^{v} I(x_k = 0, y_k = 1)$$

$$a_3 = \sum_{k=1}^{v} I(x_k = 1, y_k = 0)$$

$$a_4 = \sum_{k=1}^{v} I(x_k = y_k = 0)$$

O algoritmo de clusterização consiste em inicialmente calcular uma matriz de distâncias entre todas as unidades geográficas (municípios ou áreas mínimas comparáveis, conforme discussão na seção 3). A cada passo do processo sequencial de união de *clusters* preexistentes, a matriz de distâncias é atualizada por meio das fórmulas combinatórias (consoante seção anterior). Esta fórmulas variam por medida de dissimilaridade utilizada.

TABELA 2.1.
Expressões para as diferentes distâncias para vetores de variáveis binárias

Distância	Expressão
Jaccard	$(a_2 + a_3) / (a_1 + a_2 + a_3)$
Tanimoto	$2 \times (a_2 + a_3) / (a_1 + 2 \times (a_2 + a_3) + a_4)$
Simple mathching	$1 - (a_1 + a_4) / v$
Russel e Rao	$1 - a_1 / v$
Dice	$(a_2 + a_3) / (2 \times a_1 + a_2 + a_3)$
Kulczynski	$1 - ((a_1 / (a_1 + a_2)) + (a_1 / (a_1 + a_3))) / 2$

Elaboração dos autores.

Note-se que a medida de dissimilaridade *simple matching* no caso binário está diretamente ligada à normal euclidiana ou à norma L_1 no caso contínuo. De fato, pode-se mostrar que, se utilizarmos a distância euclidiana ou a norma L_1 para dados binários (como se eles fossem contínuos), a medida resultante é igual a v vezes a medida *simple matching*. Similarmente, a distância Tanimoto também tem uma alta correlação com a norma L_1. Estes fatos serão relevantes na análise dos resultados de comparação entre os diversos algoritmos *versus* os agrupamentos políticos de municípios, atualmente existentes (Unidades da Federação, mesorregiões e microrregiões).

Um dos problemas encontrados em Carvalho *et al.* (2009), onde se aplicam algoritmos de clusterização espacial para tratar dados contínuos, foi o fato de que, para alguns métodos de clusterização, alguns poucos *clusters* continham quase todos os municípios. Portanto, os *clusters* obtidos ao final do processo não teriam muita utilidade em termos práticos para políticas públicas, por exemplo aquelas geograficamente focadas. Para contornar este problema, os algoritmos de clusterização

foram modificados, de forma a evitar a formação de *clusters* com um número muito grande de unidades geográficas. Considerem-se as fórmulas combinatórias na seção 2.4.1, para obtenção das distâncias atualizadas $D_{J,M}$. No algoritmo modificado, quando $n_K + n_L$ são maiores do que um valor de corte, por exemplo 25, a distância entre os *clusters* J e M passa a ser $D^*_{J,M}$, onde $D^*_{J,M} = D_{J,M} + n \times$ Máx D, Máx D é a distância máxima entre todos os polígonos iniciais e n é o número de polígonos iniciais. Portanto, à nova distância é atribuído um valor inflacionado, de forma que, nos passos seguintes, não sejam formados agrupamentos com um número muito grande de unidades geográficas. Conforme discutido nas próximas seções, este artifício possibilitou a formação de *clusters* numericamente homogêneos com praticamente todos os métodos.

2.5 CRITÉRIOS PARA SELEÇÃO DO NÚMERO DE *CLUSTERS*

Na seção 2.3, apresentou-se o algoritmo sequencial de formação de novos *clusters* a partir de um conjunto de *clusters* no passo anterior. Este algoritmo continua até que haja apenas um *cluster* (ou um número mínimo de *clusters*, respeitando-se as possibilidades de vizinhança). Neste processo, podem-se construir diversos indicadores para ajudar na seleção do número de *clusters* a serem utilizados no estudo de interesse de cada pesquisador. Um dos indicadores mais populares é o critério *CCC* (*cubic clustering criterion*), de Sarle (1983), que, em algoritmos de clusterização hierárquica não espacial, testa a hipótese H_0 de que os dados foram amostrados de uma distribuição uniforme, contra a hipótese H_1 de que os dados foram amostrados de uma mistura de distribuições normais multivariadas esféricas com variâncias e probabilidades amostrais iguais. Valores positivos e altos para o *CCC* produzem a rejeição de H_0. Plotam-se os valores do *CCC* e o número de *clusters* e procura-se por picos nos quais o *CCC* excede 3. A expressão para o *CCC* é dada por

$$CCC = \log\left[\frac{1-E[R^2]}{1-R^2}\right] \times v,$$

onde v é o número de variáveis na base de dados, R^2 é o critério R^2, $E[R^2]$ é o valor esparado para o R^2 (*expected* R^2) e log[.] corresponde ao logaritmo natural. As expressões para os R^2 e o R^2 esperado são apresentadas em Sarle (1983).

Outros critérios bastante populares são o pseudo-t^2, o pseudo-F e o R^2 semiparcial. Este último mede a separação entre *clusters* no nível corrente de hierarquia. Valores altos para o pseudo-F indicam que os vetores médios de cada *cluster* são diferentes, ou seja, que cada *cluster* é significativo naquela configuração. Portanto, uma maneira de utilizar o critério pseudo-F é procurar valores de pico no gráfico da estatística pseudo-F *versus* o número de *clusters*; o número de *clusters* escolhido é o número correspondente ao pico do indicador pseudo-F. Por outro lado, se a estatística pseudo-t^2 em determinado passo da união de dois *clusters* é alta, então estes *clusters* não deveriam ser unidos, uma vez que seus vetores médios podem ser considerados diferentes. Portanto, a literatura recomenda procurar valores de picos na sequência de estatísticas pseudo-F e utilizar o número de *clusters* imediatamente superior ao número de *clusters* correspondente ao pico. Por fim, o critério R^2 semiparcial calcula a redução proporcional na variância devido à junção entre dois *clusters* (C_k e C_l). Valores pequenos indicam que os dois *clusters* podem ser considerados um só, enquanto valores altos para

o critério R^2 semiparcial indicam que os *clusters* unidos são provavelmente diferentes. Para maiores detalhes sobre os diversos critérios de escolha do número de *clusters*, consultem-se Khattree e Naik (2000).

Os critérios discutidos nesta seção referem-se especificamente a variáveis contínuas. No entanto, na ausência de outras medidas específicas para *clusters* com variáveis binárias, neste texto foram utilizados os critérios pseudo-t^2, pseudo-F, R^2 semiparcial e *CCC* para seleção do número de agrupamentos, no estudo de caso apresentado a seguir (o comportamento destes critérios para *clusters* espaciais, com variáveis contínuas, é discutido em Carvalho *et al.*, 2009). Em estudos posteriores, pretende-se analisar a possibilidade de se construírem outros critérios para seleção do número de agrupamentos especificamente para variáveis binárias.

3 ESTUDO DE CASO

Esta seção apresenta um estudo de caso para avaliar algumas das propriedades das medidas de dissimilaridade e dos tipos de distâncias apresentadas na seção 2.4. A base de dados utilizada consta de uma base de informações sobre variações positivas (um) ou negativas (zero)[2] no número total de postos de trabalhos formais existentes nos municípios brasileiros, entre 1997 e 2007, por divisão da Classificação Nacional de Atividades Econômicas – CNAE 95. A tabela A2.1, no apêndice 2, apresenta uma listagem das divisões utilizadas neste trabalho. Ao todo são 59 divisões e, portanto, foram criadas 59 variáveis binárias na base de dados. Quando se verificou crescimento do número de postos de trabalho entre 1997 e 2007, para uma determinada divisão CNAE, a esta variável atribuiu-se o valor um; quando houve um decréscimo do número de postos de trabalho entre estes dois anos, à variável atribuiu-se o valor zero. Com isto, é possível identificar, por exemplo, conglomerados de municípios onde houve aumento do número de postos de trabalho para uma grande quantidade das divisões CNAE. Os dados foram obtidos a partir da base de dados de estabelecimentos da Relação Anual de Informações Sociais (RAIS), do Ministério do Trabalho.

Além da utilização dos dados acima referidos, foi utilizada a malha de áreas mínimas comparáveis – AMCs, contendo informações georreferenciadas (vide, por exemplo, Carvalho *et al.*, 2008). Estas informações foram utilizadas na construção da estrutura de vizinhança entre os municípios. Devido à presença de ilhas no território nacional, o processo de agregação sequencial de *clusters* prosseguiu até um número mínimo de três *clusters* (quando não há mais vizinhança, seja no sentido da vizinhança do tipo *rook*, seja no sentido de vizinhança do tipo *queen*). As AMCs foram construídas para contornar a criação frequente de municípios no Brasil.

Para definir crescimento ou decréscimo do número de postos de trabalho, o critério utilizado foi o coeficiente angular β_i da seguinte regressão linear (do logaritmo natural do número de empregos):

$$\text{Log } E_{i,k,t} = \alpha_{i,k} + \beta_{i,k} \times T_t + \epsilon_{i,k,t}$$

No caso, $E_{i,k,t}$ é o número de empregos na divisão CNAE k, na AMC i, no ano t, $\alpha_{k,i}$ e $\beta_{k,i}$ são coeficientes a serem estimados via mínimos quadrados ordinários, para

2. O valor zero também foi atribuído às variações nulas.

cada AMC e para cada divisão CNAE, T_t é uma variável crescente, que corresponde a $T_t = 0$ no ano 1997 e $T_t = 10$ no ano 2007. A variável $\epsilon_{i,k,t}$ corresponde a um termo de erro da regressão linear. O exercício consiste em estimar uma regressão para cada divisão CNAE *k*, AMC *i*, obtendo-se 59 × 5.479 (vide discussão na próxima seção sobre áreas mínimas comparáveis) valores para o coeficiente $\beta_{k,i}$. Quando $\beta_{k,i} \leq 0$, à AMC *i*, para a divisão CNAE *k*, é atribuído o valor 0. Quando $\beta_{k,i} > 0$, à AMC *i*, para a divisão CNAE *k*, é atribuído o valor 1. Estas variáveis binárias, assumindo valores 0 ou 1, serão utilizadas na construção dos *clusters* espaciais.

Na seção 3.1, discute-se a compatibilização das malhas de municípios no Brasil. A seção 3.2 discute os resultados do exercício de clusterização espacial, comparando diferentes medidas de dissimilaridade. A seção 3.3 apresenta uma comparação entre os agrupamentos espaciais obtidos via clusterização espacial hierárquica e os agrupamentos políticos existentes no país; notadamente, Unidades da Federação, mesorregiões e microrregiões.

3.1 COMPATIBILIZAÇÃO DAS MALHAS DE MUNICÍPIOS

Para estudar a evolução de variáveis socioeconômicas nos municípios brasileiros, nos últimos anos, foi preciso fazer uma compatibilização da malha de municípios. Isto devido à frequente criação de novos municípios, a partir da subdivisão de municípios previamente existentes. Este estudo contou então com uma compatibilização de malhas, que permite estudar a dinâmica de variáveis locais nos anos de 1997 até 2009. Apesar de o estudo de caso apresentado aqui ter utilizado apenas variáveis entre os anos de 1997 e 2007, foi utilizada a malha criada para todo o período de 1997 a 2009 por motivo de comparabilidade com outros estudos que estão sendo concomitantemente realizados. A compatibilização de malhas municipais está baseada na ideia de áreas mínimas comparáveis – AMCs.

Entre os anos de 1997 e 2000, não houve qualquer alteração na malha de municípios. A tabela 3.1 apresenta medidas gerais sobre a evolução do número de municípios no Brasil entre 2000 e 2009. A partir de 2001 há algumas alterações na malha, com a criação de 54 novos municípios. Neste período, não foram excluídos municípios. Para compatibilização entre estas duas malhas, utilizaram-se as informações apresentadas na tabela A1.1 do Apêndice 1.

TABELA 3.1
Número de municípios e população total somada

Ano	População total	Número de munícipios
2000	169.799.170	5507
2001	172.385.776	5561
2002	174.632.932	5561
2003	176.876.251	5560
2004	179.108.134	5560
2005	184.184.074	5564
2006	186.770.613	5564
2007	189.335.191	5564
2008	189.612.814	5564
2009	191.481.045	5565

Fonte: Datasus.
Elaboração dos autores.

Entre 2001 e 2002, não foram criados nem extintos municípios. Entre 2002 e 2003, não foram criados novos municípios. Por outro lado, o município de Pinto Bandeira, código IBGE 431453, que tinha 2.673 habitantes em 2002, foi excluído. Pinto Bandeira era um distrito do município de Bento Gonçalves, no estado do Rio Grande do Sul. Foi elevado à categoria de município em 2001, mas extinto por decisão do Supremo Tribunal Federal em 2002.

Entre 2003 e 2004, não foram criados nem destruídos municípios. Entre 2004 e 2005, não foram destruídos municípios, mas foram criados quatro. A relação dos novos municípios está na tabela A1.2 no apêndice. O município de Aroeiras do Itaim é um novo município do Piauí, instalado em 1º de Janeiro de 2005, desmembramento de áreas do município de Picos. Aroeiras está localizado a 340 km da capital do estado, Teresina. O município de Figueirão é o mais novo município de Mato Grosso do Sul (no total são 78 municípios), instalado em 1º de Janeiro de 2005, desmembrado parcialmente das áreas dos municípios de Camapuã e Costa Rica (Lei nº 2.680, de 29 de setembro de 2003). O município de Ipiranga do Norte é um novo município do Mato Grosso, instalado em 1º de Janeiro de 2005, desmembramento de áreas do município de Tapurah. O município de Itanhangá é um novo município de Mato Grosso, instalado em 1º de Janeiro de 2005, desmembramento de áreas também do município de Tapurah.

Entre 2005 e 2008, não foram extintos municípios, nem criados novos municípios. Entre 2008 e 2009, foi criado o município de Nazária, com população estimada de 7.895 habitantes em 2009. Nazária foi emancipada politicamente em 1993 do município de Teresina; a sede de Nazária está a cerca de 30 quilômetros da capital. Devido a problemas jurídicos, o estatuto do novo município só foi definido em 2005, depois de decisão em última instância no Supremo Tribunal Federal. A localidade foi oficialmente instalada como município após as eleições municipais de 05 de outubro de 2008.

As áreas mínimas comparáveis, segundo o próprio nome sugere, correspondem a áreas geográficas, compostas por um ou mais municípios existentes no ano de 1997, que não foram subvididas nas criações de novos municípios até o ano de 2009. Os municípios dentro de uma mesma AMC podem ter sido subdivididos posteriormente, porém a AMC agregadamente não o foi. No total, os 5.507 municípios existentes em 1997 foram agrupados em 5.479 AMCs. Tais áreas foram utilizadas neste texto para estudar o número de postos de trabalho criados ou destruídos entre os anos de 1997 e 2007.[3]

3.2 EFEITOS DO TIPO DE DISSIMILARIDADE, DO TIPO DE DISTÂNCIA ENTRE VETORES E DO TIPO DE VIZINHANÇA ENTRE POLÍGONOS

A tabela 3.2 apresenta algumas estatísticas básicas para os tamanhos dos *clusters* formados, considerando-se a vizinhança do tipo *rook*. Para possibilitar a comparação entre os métodos, utilizaram-se 100 agrupamentos em todos os mapas. Este é o número de agrupamentos escolhido em Chein, Lemos e Assunção (2005); Carvalho *et al.* (2008), utilizam 91 agrupamentos. A tabela 3.3 apresenta as mesmas informações, referentes à vizinhança do tipo *queen*. Em ambas as tabelas, omitiu-se a

3. A lista de áreas mínimas comparáveis, compatibilizando os municípios de 1997 a 2007, pode ser fornecida pelos autores, quando requerida pelo leitor.

coluna contendo o mínimo de polígonos para cada configuração de dissimilaridade e tipo de distância; em todas as configurações, o menor *cluster* (em número de polígonos) continha exatamente um polígono. Além das tabelas apresentadas a seguir, foram também gerados diversos mapas, para cada distância, para cada método e para cada tipo de vizinhança (similarmente ao que fizeram Carvalho *et al.*, 2009). Por questão de parcimônia na apresentação dos resultados, os mapas não foram incluídos neste trabalho. A figura 3.1, discutida na seção 3.4, apresenta o mapa para agrupamentos formados com o método de Ward, e com a dissimilaridade do tipo *simple matching*.

TABELA 3.2

Comparação entre os métodos de clusterização, em termos de tamanhos dos *clusters* formados, considerando-se 100 *clusters* – vizinhança do tipo *rook*

Dissimilaridade	Distância entre vetores de variáveis binárias	Estatísticas para *clusters* com vizinhança do tipo *rook*			
		Máximo	Percentil 25%	Percentil 50%	Percentil 75%
Single linkage	Jaccard	117	35	54	70,5
	Tanimoto	112	37,5	52	69,5
	Simple matching	113	39,5	56	68,5
	Russel e Rao	152	36,5	48	66
	Dice	109	35	53	73
	Kulczynski	112	35,5	55	70
Complete linkage	Jaccard	128	41	52,5	69
	Tanimoto	129	36	50	70,5
	Simple matching	121	36	54	70,5
	Russel e Rao	154	36,5	47	70,5
	Dice	117	41	52	70,5
	Kulczynski	127	36	54	70,5
Average linkage (*unweighted*)	Jaccard	100	40	52	73
	Tanimoto	97	39	56,5	67,5
	Simple matching	92	41,5	59	67
	Russel e Rao	140	43	50	66,5
	Dice	101	38,5	52,5	73,5
	Kulczynski	100	40	53	68
Average linkage (*weighted*)	Jaccard	177	36,5	43	63
	Tanimoto	143	35	48	69
	Simple matching	133	34	42	76
	Russel e Rao	133	37	44	66,5
	Dice	171	36,5	44	60
	Kulczynski	207	37	43	74
Mediana	Jaccard	97	42	53,5	68
	Tanimoto	95	38,5	59	67
	Simple matching	103	39	58,5	66,5
	Russel e Rao	95	44	53	68,5
	Dice	93	46	53,5	66,5
	Kulczynski	105	47	53	64
Centroid	Jaccard	459	31	34	43
	Tanimoto	290	29	32	46,5
	Simple matching	392	30	32,5	46
	Russel e Rao	543	33	37	43,5
	Dice	325	31	36	46
	Kulczynski	619	32	35	41
Ward	Jaccard	150	34	45,5	72,5
	Tanimoto	145	31,5	47	72,5
	Simple matching	138	32	48,5	72,5
	Russel e Rao	240	32	42	64,5
	Dice	132	34,5	47,5	68
	Kulczynski	117	31,5	55	68

Elaboração dos autores.

Uma primeira observação importante a partir das tabelas 3.2 e 3.3 é que os maiores *clusters* contêm números moderados de polígonos. Este efeito deve-se ao artifício, utilizado neste trabalho, de penalizar a formação de *clusters* com tamanhos

muito grandes. Quando não se utiliza esta penalização, os *clusters* formados podem ser muito desiguais em termos de número de polígonos. Conforme observado em Carvalho *et al.* (2009), para *clusters* com variáveis estritamente contínuas, os métodos *single linkage*, *average linkage*, *average linkage weighted*, *centroid* e da mediana, sem penalização explícita para o tamanho dos *clusters*, tendem a formar *clusters* bastante diferentes em termos de número de unidades geográficas. Por outro lado, o método *complete linkage* e principalmente o método Ward tendem a formar *clusters* com tamanhos mais homogêneos. Portanto, a ideia de se penalizar a formação de *clusters* muito grandes, nos algoritmos de clusterização hierárquica espacial, parece ser bastante útil, pois possibilita também a utilização dos métodos *single linkage*, *average linkage*, *average linkage weighted*, *centroid* e da mediana para construção de aglomerados espaciais com tamanhos mais ou menos similares (o que pode ser mais plausível, do ponto de vista de análises geográficas).

TABELA 3.3

Comparação entre os métodos de clusterização, em termos de tamanhos dos *clusters* formados, considerando-se 100 *clusters* – vizinhança do tipo *queen*

Dissimilaridade	Distância entre vetores de variáveis binárias	Estatísticas para *clusters* com vizinhança do tipo *queen*			
		Máximo	Percentil 25%	Percentil 50%	Percentil 75%
Single linkage	Jaccard	107	35	54	72
	Tanimoto	131	40	53,5	67
	Simple matching	133	40,5	54,5	69
	Russel e Rao	151	36	47,5	69
	Dice	141	34,5	54	70
	Kulczynski	111	39,5	54,5	71,5
Complete linkage	Jaccard	133	38	51	66,5
	Tanimoto	116	39	55	68,5
	Simple matching	111	38,5	55	69
	Russel e Rao	186	36	46	64
	Dice	115	38,5	53,5	69
	Kulczynski	111	37,5	54	69,5
Average linkage (*unweighted*)	Jaccard	108	40	53	69,5
	Tanimoto	103	36	57	70
	Simple matching	88	40	57	67,5
	Russel e Rao	107	42,5	50,5	67
	Dice	105	39	54	67
	Kulczynski	114	39	54	71
Average linkage (*weighted*)	Jaccard	159	35	43	65,5
	Tanimoto	123	33,5	47,5	74,5
	Simple matching	124	32	43	76
	Russel e Rao	165	36	44,5	64,5
	Dice	150	37	44,5	67,5
	Kulczynski	156	37	43	66,5
Mediana	Jaccard	100	44	53	66,5
	Tanimoto	98	39	58	69
	Simple matching	96	38	59	67
	Russel e Rao	83	44	52,5	68
	Dice	91	43	53	68,5
	Kulczynski	80	44,5	57,5	67
Centroid	Jaccard	723	31	34	40,5
	Tanimoto	351	29	32,5	40,5
	Simple matching	389	30	32,5	39
	Russel e Rao	780	31	35	40
	Dice	403	32	35	45
	Kulczynski	675	31	35	42
Ward	Jaccard	142	34	48,5	71
	Tanimoto	129	34	47	75,5
	Simple matching	135	33,5	49	69
	Russel e Rao	183	32,5	42	64,5
	Dice	144	34,5	47,5	72
	Kulczynski	148	33	49,5	67

Elaboração dos autores.

Em termos de tamanho máximo dos *clusters*, para os diferentes métodos e as diferentes distâncias, o método da mediana apresenta os *clusters* com menores tamanhos máximos. O método *centroid*, por sua vez, incorre em *clusters* mais desiguais em termos de tamanho. Em relação aos tipos de distância entre vetores com dados binários, a distância do tipo *simple matching* incorre na formação de *clusters* com tamanhos menos desiguais, enquanto a distância do tipo Russel e Rao gera *clusters* com tamanhos mais dispersos.

A partir das tabelas 3.2 e 3.3, e dos mapas gerados, notou-se que os resultados utilizando-se a vizinhança do tipo *rook* são bem semelhantes aos resultados utilizando-se a vizinhança do tipo *queen*, em se tratando da distribuição do número de municípios em cada *cluster*. Observou-se também que os agrupamentos formados com os dois tipos de vizinhança podem ser bastante diferentes, ainda que utilizando-se o mesmo método de clusterização e a mesma distância entre vetores.

Em princípio, a utilização da vizinhança do tipo *queen*, por exigir apenas um vértice em comum para caracterizar vizinhança, pode implicar na formação de agrupamentos mais irregulares do que os formados pela vizinhança do tipo *rook*. Por outro lado, por se tratar de um tipo de contiguidade menos restritivo, espera-se que os *clusters* com vizinhança do tipo *queen* apresentem menor variabilidade *intraclusters*. O exercício apresentado na próxima seção aborda mais diretamente este tópico.

3.3 COMPARAÇÃO COM AGRUPAMENTOS POLÍTICOS DE MUNICÍPIOS BRASILEIROS

Nesta seção, apresenta-se uma comparação dos agrupamentos obtidos via clusterização espacial hierárquica *versus* os agrupamentos políticos de municípios existentes no Brasil. Os agrupamentos utilizados, para fins comparativos, são: microrregiões, mesorregiões e Unidades da Federação.[4] Ao todo, perfazem 27 Unidades da Federação, 558 microrregiões e 137 mesorregiões. Portanto, para comparar os resultados dos *clusters* às divisões políticas, utilizaram-se configurações com 27, 558 e 137 agrupamentos. Para cada um destes três números de agrupamentos, calculou-se a soma dos quadrados dos desvios em relação à média de cada *cluster* (*TCSS*). Esta foi a medida utilizada enquanto indicador de performance[5] da configuração de agrupamentos – ela fornece uma ideia da variabilidade *intraclusters* para cada método. A expressão para o *TCSS* é dada por:

$$TCSS = \sum_{k=1}^{G} \sum_{i \in C_k} \sum_{l=1}^{v} [x_{k,i,l} - \bar{x}_{k,l}]^2 ,$$

onde v é o número total de variáveis na base de dados, G é o número de *clusters* (G = 27, 137 ou 558), C_k é o conjunto de municípios no *cluster k*, $x_{k,i,l}$ é a variável l no município i, e $\bar{x}_{k,l}$ é a média da variável l, dentro do *cluster k*. Para comparação

4. A comparação com Unidades da Federação tem caráter mais ilustrativo, uma vez que os critérios para definição dos estados brasileiros, em muitos casos, foram pautados mais por critérios histórico-políticos, do que por critérios de homogeneidade regional de fato.

5. Apesar de as variáveis na base de dados serem estritamente binárias (zero ou um), utilizou-se essa medida de performance devido à sua simplicidade e facilidade de interpretação. Outras medidas de desempenho podem ser empregadas especificamente para tratamento de variáveis binárias.

entre os diversos métodos de clusterização e as divisões políticas, os valores a serem reportados são a variabilidade relativa de cada método *versus* a variabilidade da divisão política correspondente. Neste caso, a variabilidade relativa $\Delta TCSS_{Método}$ é dada por:

$$\Delta TCSS_{Método} = 100 \times TCSS_{Método} / TCSS_{Divisão\ Política}$$

A tabela 3.4 a seguir apresenta a medida de performance para diferentes tipos de medidas de dissimilaridade e diferentes distâncias entre vetores. As últimas três linhas na tabela 3.4 servem apenas para explicitar que as medidas de avaliação dos *clusters* formados são relativas aos agrupamentos políticos existentes, para os quais o valor básico é 100. Note-se que os *clusters* obtidos através de clusterização espacial hierárquica, para o método Ward, apresentam menores variabilidades totais do que os agrupamentos políticos. Esta observação vale para microrregiões (558 *clusters*), mesorregiões (137 *clusters*) e Unidades da Federação (27 *clusters*). No caso de microrregiões, o método de Ward implica uma redução de variabilidade de até quase 17%. Comparando-se aos resultados encontrados em Carvalho *et al.* (2009), os resultados da tabela 3.4 mostram que, devido ao artifício de penalização de *clusters* muito grandes, todas as medidas de dissimilaridade e todos os métodos de clusterização resultaram em redução na variabilidade total, quando comparados às microrregiões.[6] Para as mesorregiões, os resultados mostram que todos os métodos, combinados com as diferentes medidas de dissimilaridade, também incorreram em medidas de variabilidade *intraclusters* menores do que as medidas para os agrupamentos políticos. Finalmente, para o caso de Unidades de Federação, os agrupamentos numéricos (via algoritmos hierárquicos) não mais foram superiores aos agrupamentos políticos. Isto pode ser explicado devido ao fato de as Unidades da Federação já serem agrupamentos muito grandes.

O fato de os diversos métodos de clusterização, principalmente o método de Ward, terem gerado *clusters* com menor variabilidade do que as divisões políticas não significa que os *clusters* obtidos via algoritmos numéricos são superiores aos agrupamentos políticos já existentes. A ideia, no entanto, é que, para estudos nos quais o objetivo do pesquisador seja utilizar medidas de agrupamentos os mais homogêneos possíveis, a utilização de agrupamentos formados via clusterização hierárquica espacial pode ser mais adequada, no estudo específico, do que a utilização de divisões políticas já existentes. Além disso, os agrupamentos podem ser gerados de acordo com um conjunto de variáveis específicas de interesse do pesquisador.

O método de clusterização hierárquica que resultou na menor medida de variabilidade foi o método de Ward, utilizando as medidas de dissimilaridade *simple matching* e Tanimoto. Isto já era esperado, pois, conforme reportado na literatura de clusterização hierárquica tradicional (não hierárquica), o algoritmo de Ward tende a

6. Conforme Carvalho *et al.* (2009), para os métodos *single linkage*, *average linkage*, *average linkage weighted*, da mediana e *centroid*, as medidas de variabilidade *intraclusters* são maiores do que a variabilidade dada pelas divisões políticas, no caso de Unidades da Federação e mesorregiões. Para microrregiões, os métodos *single linkage*, da mediana e *centroid* apresentaram variabilidades maiores do que as divisões políticas. Esta maior variabilidade para tais métodos está diretamente ligada ao fato de que alguns dos *clusters* formados continham quase todos os municípios no território nacional. Este problema foi resolvido mediante utilização da penalização para a formação de *clusters* muito grandes.

formar *clusters* de forma que a variância agregada seja minimizada. Além disso, a medida de variabilidade (*TCSS*) utilizada foi construída usando-se intrinsecamente a norma euclidiana, e esta está intrinsecamente relacionada à dissimilidade *simple matching* e à dissimilaridade de Tanimoto. Possivelmente, utilizando-se outras medidas de variabilidade, mais ligadas a outros tipos de dissimilaridade, a variabilidade resultante seria menor no caso de métodos de clusterização baseados na dissimilaridade de Kulczynski, por exemplo. Em todo caso, o presente exercício reforça a utilização dos algoritmos de clusterização hierárquica espacial. A seleção do melhor método e da melhor distância, uma vez que o artifício de penalização de *clusters* muito grandes permitiu a construção de *clusters* geográficos com tamanhos mais homogêneos, vai depender de critérios, entre eles visuais, de inspeção dos mapas formados. Métodos mais objetivos de seleção do melhor método, da melhor distância, da melhor vizinhança e do número de agrupamentos utilizados estão atualmente sob investigação pelos autores.

Vale notar que, por permitir maior flexibilização na formação dos *clusters*, a vizinhança do tipo *queen* implica em agrupamentos com variabilidade um pouco menor do que os agrupamentos obtidos com a vizinhança do tipo *rook*. Isto era esperado, dado o caráter menos restritivo da vizinhança do tipo *queen*; mais pares de municípios são considerados vizinhos, uma vez que apenas um vértice em comum é requerido para definição de vizinhança. Em todo caso, se, ao invés de um algoritmo de natureza hierárquica, estivéssemos utilizando um algoritmo de minimização explícita da variabilidade total, possivelmente a vizinhança do tipo *queen* incorreria em variabilidade ainda menor do que a vizinhança do tipo *rook*. No entanto, o algoritmo hierárquico não é um algoritmo de minimização explícita, de forma que os *clusters* formados não necessariamente correspondem aos *clusters* formados por um algoritmo de maximização de alguma função objetivo.

TABELA 3.4
Percentual da variabilidade total dos diferentes métodos de clusterização, em comparação com as divisões políticas de municípios brasileiros

Metodologia de clusterização hierárquica espacial		Número de agrupamentos					
Dissimilaridade	Distância entre vetores de variáveis binárias	27		137		558	
		Rook	Queen	Rook	Queen	Rook	Queen
Single linkage	Jaccard	101,1	101,0	98,5	98,7	99,2	98,1
	Tanimoto	101,1	101,4	98,0	97,1	91,5	90,8
	Simple matching	101,0	100,1	98,4	96,8	91,5	90,9
	Russel e Rao	100,4	102,0	99,2	98,9	97,7	97,5
	Dice	100,7	101,6	98,6	98,2	99,3	98,1
	Kulczynski	101,3	101,1	99,0	98,9	98,7	98,6
Complete linkage	Jaccard	100,7	101,6	96,3	96,1	91,9	91,7
	Tanimoto	100,7	100,3	96,0	96,0	85,4	85,1
	Simple matching	99,9	101,1	95,8	95,7	85,4	85,0
	Russel e Rao	101,0	101,2	96,4	96,9	94,5	94,4
	Dice	100,3	100,7	96,2	95,9	91,6	91,6
	Kulczynski	101,5	101,2	98,3	98,5	96,0	95,6

(Continua)

(Continuação)

Metodologia de clusterização hierárquica espacial		Número de agrupamentos					
Dissimilaridade	Distância entre vetores de variáveis binárias	27		137		558	
		Rook	Queen	Rook	Queen	Rook	Queen
Average linkage (*unweighted*)	Jaccard	100,8	100,5	97,6	97,0	94,5	94,2
	Tanimoto	100,8	100,6	97,9	97,1	89,4	88,9
	Simple matching	101,0	100,6	97,3	97,1	88,9	88,3
	Russel e Rao	100,8	101,3	96,5	96,8	92,2	91,9
	Dice	100,7	100,1	97,2	96,8	94,1	93,6
	Kulczynski	100,9	101,8	98,2	97,2	96,2	95,4
Average linkage (*weighted*)	Jaccard	100,2	99,9	96,0	95,3	92,6	92,6
	Tanimoto	100,6	100,1	96,9	96,4	88,1	87,5
	Simple matching	100,4	100,8	96,7	96,5	87,4	87,2
	Russel e Rao	100,5	100,6	95,6	95,9	91,8	91,4
	Dice	99,7	100,1	96,2	95,5	92,9	92,8
	Kulczynski	100,6	100,8	96,2	96,3	94,2	93,7
Mediana	Jaccard	101,0	101,1	98,2	98,0	97,0	97,1
	Tanimoto	101,0	101,0	98,3	98,1	91,9	91,5
	Simple matching	100,6	100,2	98,1	98,1	91,9	91,8
	Russel e Rao	101,4	100,8	97,2	97,3	92,7	92,1
	Dice	100,8	101,0	99,5	98,4	97,3	97,6
	Kulczynski	100,4	100,4	98,7	98,4	98,1	98,0
Centroid	Jaccard	102,8	103,4	98,0	97,1	94,7	93,9
	Tanimoto	101,8	101,5	97,8	97,6	90,0	89,6
	Simple matching	102,1	101,5	98,2	97,4	89,9	89,6
	Russel e Rao	102,4	101,5	98,2	97,2	92,9	92,1
	Dice	101,9	100,9	97,6	97,6	94,9	94,2
	Kulczynski	101,8	102,1	98,7	97,7	96,2	96,2
Ward	Jaccard	98,3	98,6	96,1	95,3	88,4	87,8
	Tanimoto	98,8	98,6	95,5	94,5	83,7	83,1
	Simple matching	98,9	99,0	95,1	95,0	83,6	83,1
	Russel e Rao	96,4	95,0	93,4	92,7	88,9	88,8
	Dice	98,0	97,7	94,9	94,8	88,4	88,2
	Kulczynski	98,8	98,6	96,0	96,5	91,0	90,7
Divisão política de municípios	Unidades da Federação		100		---		---
	Mesorregiões		---		100		---
	Microrregiões		---		---		100

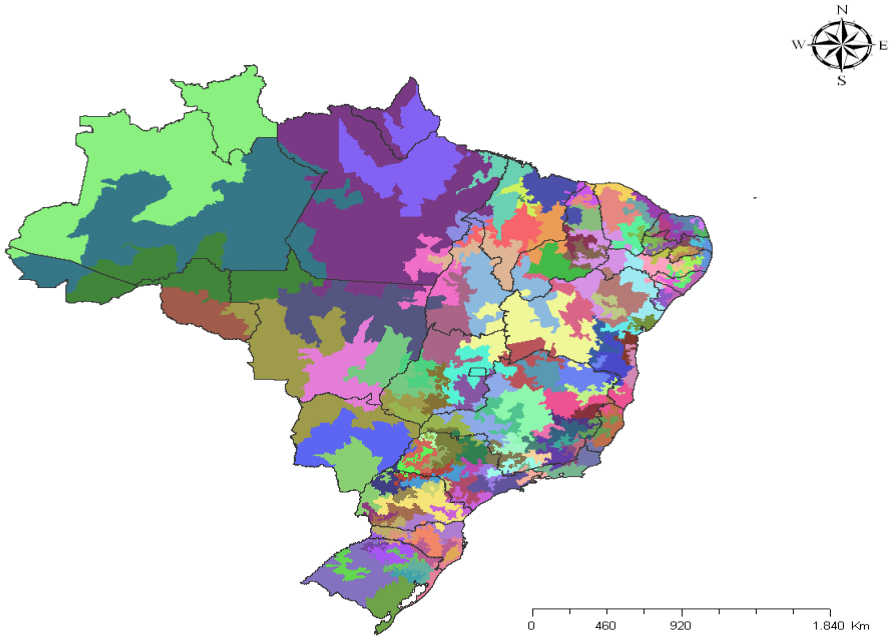
Elaboração dos autores.

3.4 DETECÇÃO DE ÁREAS DE PROSPERIDADE COM DIVERSIFICAÇÃO

No estudo apresentado nesta seção, construíram-se agrupamentos homogêneos com base nas variações positivas e negativas do número de postos de trabalho formais nos municípios brasileiros, entre os anos de 1997 e 2007. Foram consideradas variações por divisão da classificação CNAE 95. Uma vez construídos os agrupamentos, é possível identificar aqueles que foram mais prósperos, com diversificação no seu crescimento, neste intervalo de dez anos. Para isso, consideraram-se os *clusters* obtidos com o método Ward e a distância *simple matching*. O número de agrupamentos

escolhido foi igual a 133, com base nos critérios de seleção do número de agrupamentos (pseudo-*F*, *CCC*, R^2 semiparcial, pseudo-t^2). O mapa com os 133 agrupamentos gerados encontra-se na figura 3.1 (escolheram-se este método de clusterização e esta distância por terem apresentado menores medidas de variabiliadade no exercício da seção anterior).

FIGURA 3.1
Clusters* espaciais – método Ward, distância *simple matching*, vizinhança *rook



Elaboração dos autores.
Obs.: O mapa está reproduzido conforme o original fornecido pelos autores, cujas características não permitiram melhor ajuste para fim de impressão (nota do Editorial).

Para identificar as áreas de maior prosperidade, considere-se o *cluster* C_k. Para cada variável (divisão CNAE) l, calcula-se a média $\bar{x}_{k,l}$, dentro do *cluster* C_k. Esta média corresponde à proporção de municípios dentro do *cluster* C_k que tiveram crescimento médio no número de empregos formais na divisão CNAE l. Portanto, *clusters* com altas médias $\bar{x}_{k,l}$, para um grande número de divisões CNAE, indicam áreas de crescimento da economia, de forma diversificada. A medida de prosperidade diversificada MPD_k para o *cluster* C_k tem expressão:

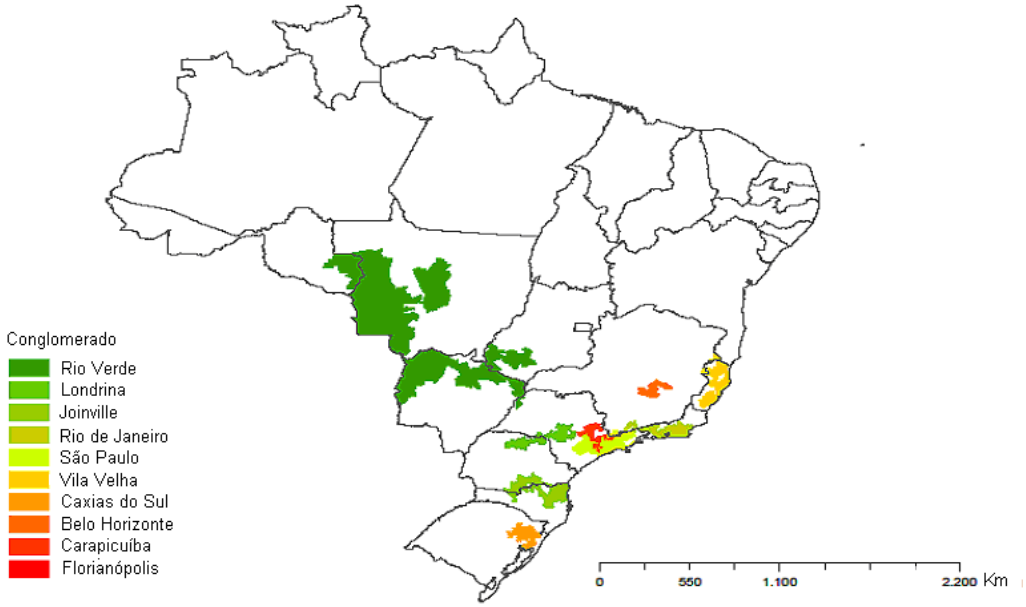
$$MPD_k = 100 \times \sum_{l=1}^{v} \bar{x}_{k,l} \ .$$

Portanto, quanto mais setores apresentarem crescimento médio positivo no *cluster* C_k, maior a medida MPD_k para este *cluster*. A figura 3.2 apresenta a localização das dez áreas com maior coeficiente MPD_k.

A tabela 3.5 a seguir apresenta a caracterização dos dez agrupamentos com maiores medidas de prosperidade com diversificação. A primeira coluna apresenta o nome dado ao agrupamento – este nome refere-se ao município de maior população dentro do agrupamento. A segunda coluna na tabela 3.5 apresenta a lista dos municípios em cada agrupamento. A terceira coluna mostra a população total em

2009 (soma das populações de cada município) em cada agrupamento, e a quarta coluna apresenta a medida (MPD) para cada agrupamento.

FIGURA 3.2
Áreas com crescimento diversificado, entre 1997 e 2007



Elaboração dos autores.
Obs.: O mapa está reproduzido conforme o original fornecido pelos autores, cujas características não permitiram melhor ajuste para fim de impressão (nota do Editorial).

O município de Florinópolis, que compõe um *cluster* sozinho, foi a área que apresentou maior crescimento com diversificação. Neste município, observou-se crescimento líquido em praticiamente três quartos das divisões CNAE. O fato de o *cluster* composto por Florianópolis ter apresentado maior crescimento não significa necessariamente que Florianópolis, como capital, teve desempenho melhor do que as demais capitais do país. O que os resultados da clusterização mostraram é que o *cluster* formado apenas por Florianópolis[7] apresentou desempenho melhor (considerando-se a medida de desempenho utilizada) do que os demais *clusters*, que continham as demais capitais brasileiras.

A segunda área com maior crescimento diversificado foi o agrupamento contendo o município de Carapicuíba. O desempenho deste *cluster* pode ser explicado pelo crescimento do entorno da região metropolitana de São Paulo, que transbordou para municípios um pouco mais afastados. Pelo fato de estes municípios serem menores, no início do período considerado (1997), do que a região metropolitana propriamente dita, o crescimento percentual observado foi maior do que nos demais *clusters*. O agrupamento contendo o município de Belo Horizonte foi o terceiro agrupamento em termos de crescimento diversificado. Em geral, espera-se que a medida MPD seja

7. O *cluster* contendo apenas Florianópolis decorreu da diferenciação desta cidade em relação aos seus vizinhos geográficos. Portanto, durante os passos de formação da árvore de *clusters*, até o ponto de parada do algoritmo, Florianópolis não se juntou aos vizinhos.

maior nos municípios onde já existe atividade econômica diversificada. Em grande parte dos municípios brasileiros, muitas das 59 divisões CNAE não possuem estabelecimentos formais, de forma que não se observa crescimento ou decrescimento naquela área – simplesmente aquele tipo de atividade não existe no município em termos formais. Foi este o motivo que levou os autores a não fazerem um índice equivalente para identificação de municípios com baixo crescimento ou com crescimento sem diversificação. Pela existência de muitos municípios com pouca diversificação de atividade econômica estruturalmente no período estudado, os resultados de uma medida de descréscimo agregado teriam difícil interpretação.

TABELA 3.5
Caracterização dos dez agrupamentos que apresentaram crescimento com prosperidade, entre 1997 e 2007

Cluster	Municípios	População Total	Medida de Prosperidade Diversificada (MPD)
Florianópolis	Florianópolis	408.163	74,58%
Carapicuíba	Carapicuíba, Piracicaba, Jundiaí, Limeira, Barueri, Taboão da Serra, Itapevi, Americana, Rio Claro, Santa Bárbara d'Oeste, Indaiatuba, Cotia, Bragança Paulista, Mogi Guaçu, Atibaia, Araras, Santana de Parnaíba, Jandira, Salto, Valinhos, Várzea Paulista, Itatiba, Leme, Itapira, São Roque, Vinhedo, Nova Odessa, Monte Mor, Vargem Grande Paulista, Cabreúva, Itupeva, Louveira, Conchal, Piracaia, Santa Gertrudes, Cordeirópolis, Iracemápolis, São Lourenço da Serra, Araçariguama	4.917.982	67,01%
Belo Horizonte	Belo Horizonte, Contagem, Betim, Ribeirão das Neves, Santa Luzia, Ibirité, Sabará, Itabira, Vespasiano, Itaúna, Pará de Minas, Nova Lima, Pedro Leopoldo, Lagoa Santa, Itabirito, Matozinhos, Brumadinho, Igarapé, Mateus Leme, Sarzedo, São Joaquim de Bicas, Juatuba, São José da Lapa, Nova Era, Jaboticatubas, Mário Campos, Florestal, Confins	5.273.781	57,63%
Caxias do Sul	Caxias do Sul, Viamão, Bento Gonçalves, Farroupilha, Parobé, Estância Velha, Roca Sales, Garibaldi, Canela, Santo Antônio da Patrulha, Gramado, Igrejinha, Imigrante – Teutônia, Portão, Flores da Cunha, Dois Irmãos, Veranópolis, Carlos Barbosa, Três Coroas, São Francisco de Paula, São Sebastião do Caí, São Marcos, Encantado, Rolante, Ivoti, Nova Petrópolis, Nova Hartz, Antônio Prado, Feliz, Capela de Santana, Glorinha, Ipê, Araricá, Nova Roma do Sul, Nova Pádua	1.546.269	56,58%
Vila Velha	Vila Velha, Serra, Cariacica, Vitória, Cachoeiro de Itapemirim, Linhares, Colatina, Guarapari, São Mateus, Aracruz, Viana, Nova Venécia, Barra de São Francisco, Nanuque, Santa Maria de Jetibá, Castelo, Domingos Martins, São Gabriel da Palha, Mantena, Pedro Canário, Pinheiros, Anchieta, Venda Nova do Imigrante, Montanha, Vargem Alta, Piúma, Fundão, Alfredo Chaves, Marechal Floriano, Iconha, Rio Novo do Sul, Ibiraçu, Marilândia, Atilio Vivacqua	2.834.310	54,79%
São Paulo	São Paulo, Guarulhos, São Bernardo do Campo, Osasco, Santo André, São José dos Campos, Sorocaba, Mauá, Santos, Diadema, Mogi das Cruzes, Itaquaquecetuba, São Vicente, Guarujá, Suzano, Taubaté, Praia Grande, Embu, Jacareí, Ferraz de Vasconcelos, Itapecerica da Serra, Itu, Francisco Morato, São Caetano do Sul, Itapetininga, Pindamonhangaba, Franco da Rocha, Cubatão, Guaratinguetá, Poá, Ribeirão Pires, Tatuí, Votorantim, Caraguatatuba, Caieiras, Itanhaém, Caçapava, Lorena, Ubatuba, Arujá, Mairiporã, Campo Limpo Paulista, São Sebastião, Ibiúna, Cajamar, Embu-Guaçu, Piedade, Porto Feliz, Santa Isabel, Campos do Jordão, Capivari, Boituva, Bertioga, Mongaguá, Mairinque, Rio Grande da Serra, Tremembé, Salto de Pirapora, Cerquilho, Aparecida, Tietê, São Miguel Arcanjo, Biritiba-Mirim, Juquitiba, Rio das Pedras, Pilar do Sul, Iperó, Extrema, Guararema, Araçoiaba da Serra, Cambuí, Jarinu, Potim, Camanducaia, Bom Jesus dos Perdões, Capela do Alto, Paraibuna, Alumínio, Salesópolis, Pirapora do Bom Jesus, Elias Fausto, Nazaré Paulista, Santa Branca, Joanópolis, Roseira, Igaratá, Sarapuí, Rafard, Saltinho, Vargem, Santo Antônio do Pinhal, Sapucaí-Mirim, Jambeiro, Alambari, Monteiro Lobato, Mombuca	23.781.735	53,97%
Rio de Janeiro	Rio de Janeiro, Nova Iguaçu, São Gonçalo, Duque de Caxias, Belford Roxo, Niterói, São João de Meriti, Petrópolis, Volta Redonda, Magé, Itaboraí, Macaé, Cabo Frio, Nova Friburgo, Barra Mansa, Teresópolis, Nilópolis, Queimados, Resende, Maricá, Araruama, Itaguaí, Barra do Piraí, Japeri, Rio das Ostras, Itajubá, São Pedro da Aldeia, Cruzeiro, Seropédica, Valença, Saquarema, Cachoeiras de Macacu, Rio Bonito, Guapimirim, Paracambi, São Lourenço, Casimiro de Abreu, Tanguá, Armação dos Búzios, Arraial do Cabo, Bom Jardim, Piraí, Miguel Pereira, Iguaba, Grande, Pinheiral, Silva Jardim, Caxambu, Porto Real, Passa Quatro, Itanhandu, Maria da Fé, Carmo de Minas, Quatis, Cristina, Duas Barras, Pouso Alto	14.705.910	53,75%

(Continua)

(Continuação)

Cluster	Municípios	População Total	Medida de Prosperidade Diversificada (MPD)
Joinville	Joinville, Blumenau, Itajaí, Jaraguá do Sul, Brusque, Balneário Camboriú, São Bento do Sul, Rio do Sul, Camboriú, Navegantes, Gaspar, Canoinhas, União da Vitória, Indaial, Rio Negrinho, Palmas, São Mateus do Sul, São Francisco do Sul, Itapema, Timbó, Porto União, Guaramirim, Rio Negro, Pomerode, Joaçaba, Araquari, Penha, Ituporanga, Barra Velha, Cruz Machado, Três Barras, Papanduva, Ibirama, Bituruna, General Carneiro, Balneário Piçarras, Rebouças, Massaranduba, São João do Triunfo, Presidente Getúlio, Garuva, Schroeder, Rio Azul, Corupá, Mallet, Ilhota, Campo Alegre, Itapoá, Inácio Martins, Rodeio, Apiúna, Benedito Novo, Rio dos Cedros, Agrolândia, Lontras, Luiz Alves, Santa Terezinha, Balneário Barra do Sul, Antônio Olinto, Rio do Oeste, Água Doce, Ascurra, Petrolândia, Laurentino, Vitor Meireles, Fernandes Pinheiro, Luzerna, José Boiteux, Agronômica, Porto Vitória, Witmarsum, Dona Emma, Doutor Pedrinho, Braço do Trombudo, São João do Itaperiú, Atalanta	2.728.605	50,56%
Londrina	Londrina, Bauru,Maringá, Jaú, Botucatu, Apucarana, Ourinhos, Arapongas, Cambé, Sarandi, Avaré, Lençóis Paulista, Rolândia, Cornélio Procópio, Ibiporã, Santa Cruz do Rio Pardo, Pederneiras, Santo Antônio da Platina, Jacarezinho, Barra Bonita, Agudos, Mandaguari, Bandeirantes, Marialva, Astorga, Cambará, Palmital, Andirá, Cerqueira César, Macatuba, Assaí, Sertanópolis, Uraí, Jataizinho, Manduri, Arandu, Iaras, Sabáudia, Águas de Santa Bárbara, Barra do Jacaré, Óleo, Borebi	2.846.312	50,12%
Rio Verde	Rio Verde, Cáceres, Lambari D'Oeste, Mirassol d'Oeste, Sinop, Corumbá, Jataí, Tangará da Serra, Cacoal, Vilhena, Sorriso, Andradina, Rolim de Moura, Mineiros, Pontes e Lacerda, Paranaíba, Juína, Camapuã, Costa Rica, Barra do Bugres, Pimenta Bueno, Lucas do Rio Verde, Coxim, Nova Mutum, Espigão D'Oeste, Ilha Solteira, Campo Novo do Parecis, Cassilândia, São Gabriel do Oeste, Tapurah, Aparecida do Taboado, São José dos Quatro Marcos, Diamantino, Comodoro, Ladário, Colorado do Oeste, Chapadão do Sul, Araputanga, Sapezal, Nobres, Vila Bela da Santíssima Trindade, Jauru, Porto Esperidião, Vera, Chapadão do Céu, Nova Lacerda, Campos de Júlio, Rio Branco, Figueirópolis D'Oeste, Salto do Céu, Portelândia, Glória D'Oeste, Indiavaí, Reserva do Cabaçal	1.764.674	50,05%
Total	---	60.807.741	---

Elaboração dos autores.

4 CONCLUSÕES

Neste trabalho estudou-se uma metodologia para formação hierárquica de agrupamentos espaciais, para o caso no qual as variáveis são estritamente binárias (zero ou um). O algoritmo estudado é uma modificação do algoritmo aglomerativo hierárquico de clusterização tradicional: a cada passo do processo de junção de dois *clusters*, para formação de um novo, impõe-se que a junção pode acontecer somente entre *clusters* geograficamente vizinhos (de acordo com um sistema de dados georreferenciados). Neste caso, consideraram-se dois tipos de vizinhanças: vizinhança do tipo *rook* (polígonos com uma aresta em comum); e vizinhança do tipo *queen* (polígonos com um vértice em comum). Devido ao fato de o algoritmo de clusterização espacial hierárquica estudado neste trabalho ser uma extensão do algoritmo de clusterização hierárquica tradicional, pode-se importar os tipos de dissimilaridade empregados na literatura conhecida. Os tipos de dissimilaridade ou métodos empregados neste texto são: *Ward minimum variance*, *centroid*, mediana, *single linkage*, *complete linkage*, *average linkage*, *average linkage weigthed*. Além disso, empregaram-se diferentes definições de distâncias entre vetores de variáveis binárias. As distâncias empregadas são: Jaccard, Tanimoto, *simple matching*, Russel e Rao, Dice, Kulczynski.

Este estudo é o segundo passo em um projeto de pesquisa conduzido pelos autores para construção de algoritmos de clusterização espacial. Em outro trabalho, Carvalho *et al.* (2009), os autores estudaram a formação de *clusters* espaciais com variáveis estritamente contínuas. No caso, os resultados mostraram que os métodos de

Ward e *complete linkage* tendem a fornecer *clusters* com tamanhos não tão desiguais. Por outro lado, os demais métodos tendem a formar *clusters* com tamanhos bastante diferentes. Por este motivo, no presente texto introduziu-se um artifício para evitar a formação de agrupamentos muito heterogêneos em termos de número de polígonos. A cada passo do processo de formação de novos agrupamentos a partir da função de agrupamentos anteriores, introduz-se um pênalti para a formação de *cluster*s com número de componentes acima de um valor de corte. Os resultados aqui apresentados utilizaram um corte igual a 25.

Este texto apresenta um estudo de caso para avaliar algumas das propriedades das medidas de dissimilaridade e dos tipos de distâncias para vetores de variáveis binárias. A base de dados utilizada refere-se a uma base de informações sobre variações positivas (um) ou negativas (zero) no número total de postos de trabalhos formais existentes nos municípios brasileiros, entre 1997 e 2007, por divisão da Classificação Nacional de Atividades Econômicas – CNAE 95. Ao todo são 59 divisões e, portanto, foram criadas 59 variáveis binárias na base de dados. Quando houve um crescimento do número de postos de trabalho entre 1997 e 2007, para uma determinada divisão CNAE, a esta variável se atribuiu o valor um; quando houve um decréscimo do número de postos de trabalho entre estes dois anos, à variável foi atribuído o valor zero. Com isso, é possível identificar, por exemplo, conglomerados de municípios onde houve aumento do número de postos de trabalho para uma grande quantidade das divisões CNAE. Os dados foram obtidos a partir da base de dados de estabelecimentos RAIS, do Ministério do Trabalho. Entre os resultados obtidos, notou-se que a penalização da formação de *clusters* com tamanho acima de um valor de corte permitiu a formação de *clusters* com tamanhos mais homogêneos, para todos os métodos. Isto aumenta o número de opções utilizáveis para a formação de agrupamentos espaciais.

O texto traz também uma comparação dos agrupamentos obtidos via clusterização espacial hierárquica com agrupamentos políticos de municípios existentes no Brasil. Os agrupamentos políticos utilizados foram: microrregiões, mesorregiões e Unidades da Federação. Ao todo, 27 Unidades da Federação, 558 microrregiões e 137 mesorregiões. Portanto, para comparar os resultados dos *clusters* às divisões políticas, utilizaram-se configurações com 27, 558 e 137 agrupamentos. Para cada um destes três números de agrupamentos, calculou-se a soma dos quadrados dos desvios em relação à média de cada *cluster*, enquanto medida de variabilidade total *intraclusters*. Os resultados mostraram a capacidade de todos os métodos, principalmente do método de Ward, em gerar agrupamentos com variabilidade menor do que os agrupamentos políticos. No caso de microrregiões, por exemplo, o método de Ward possibilitou a formação de agrupamentos homogêneos, com 17% a menos da variabilidade dos agrupamentos políticos. Para os demais métodos, devido ao artifício de penalização da formação de *clusters* com tamanhos acima de um valor de corte, a variabilidade total obtida resultou, na maioria das situações, menor do que a variabilidade no caso dos agrupamentos políticos. Para microrregiões, todos os métodos utilizados, e todas as medidas de distância utilizadas, resultaram em redução na variabilidade total.

Comparando-se os resultados obtidos com os dois tipos de vizinhança, em termos de visualização geográfica, a vizinhança do tipo *queen* permite uma maior flexibilidade na formação dos agrupamentos, uma vez que exige apenas um vértice em comum para caracterizar vizinhança entre dois municípios. A vizinhança do tipo *rook*, por exigir um vértice em comum para caracterizar vizinhança, possibilita a formação de agrupamentos com forma menos irregular. Devido ao fato de a vizinhança do tipo *queen* ser menos restritiva, a medida de variabilidade *intraclusters* utilizando-se a vizinhança *queen* resultou um pouco menor do que a medida de variabilidade utilizando-se a vizinhança do tipo *rook*. Esta diferença poderia ser maior, caso estivesse sendo utilizado um algoritmo de optimização estrita de uma função objetivo global, ao invés de um algoritmo aglomerativo de natureza hierárquica.

Finalmente, com base nos algoritmos investigados neste estudo, construíram-se agrupamentos homogêneos com base nas variações positivas e negativas do número de postos de trabalho formais nos municípios brasileiros, entre os anos de 1997 e 2007. Foram consideradas variações por divisão da classificação CNAE 95. Uma vez construídos os agrupamentos, é possível identificar aqueles que foram mais prósperos, com diversificação no seu crescimento, neste intervalo de dez anos. Para isso, consideraram-se os *clusters* obtidos com o método de Ward e a distância *simple matching*. O número de agrupamentos escolhido foi igual a 133, com base nos critérios de seleção do número de agrupamentos (pseudo-*F*, *CCC*, R^2 semiparcial, pseudo-t^2). Para identificar as áreas de maior prosperidade, criou-se uma medida de prosperidade diversificada MPD_k para cada *cluster* C_k. Quanto mais setores apresentarem crescimento médio positivo no *cluster* C_k, maior a medida MPD_k para este *cluster*. Florianópolis foi o município com maior medida *MPD*, e todas as dez áreas com maiores valores para este coeficiente estão localizadas nas regiões Sudeste, Sul e Centro-Oeste do país.

Várias questões permanecem em aberto para futuras investigações, o que poderá ensejar o aprimoramento da metodologia aqui estudada. Primeiramente, seria interessante estender as distâncias entre vetores para incorporar variáveis categóricas nominais e ordinais (com mais de duas categorias). Além disso, seria interessante também desenvolver metodologias específicas para tratar da combinação de diferentes tipos de variáveis (categóricas e contínuas, por exemplo), em uma mesma base de dados (Hiu *et al.*, 2001, apresentam um algoritmo de clusterização com diferentes tipos de dados; o algoritmo utilizado não se baseia em procedimentos hierárquicos).

Os algoritmos estudados neste texto são puramentente heurísticos, e não estão baseados em modelos probabilíticos intrínsecos, a partir dos quais procedimentos de estimação bayesiana ou métodos de máxima verossimilhança podem ser utilizados. Mesmo os procedimentos heurísticos aqui apresentados não necessariamente seguem os mesmos comportamentos teóricos estudados em artigos sobre clusterização hierárquica não espacial. Abre-se então uma frente de estudos para outros métodos de clusterização baseados em modelos probabilísticos, e outra frente para avaliar formalmente as propriedades dos procedimentos heurísticos aqui estudados.

Finalmente, outra abertura para novos estudos é a seleção do número de *clusters*. Apresentou-se aqui uma discussão sobre métodos tradicionais de seleção do número de *clusters*. No entanto, as propriedades levantadas para estes critérios, no caso de algoritmos não espaciais, muito provavelmente não se aplicam aos métodos heurísticos de clusterização espacial. De fato, um ponto importante é que, em clusterização não espacial, muitas vezes o interesse está em se chegar a um número razoavelmente pequeno de agrupamentos (da ordem de 10); tendo poucos agrupamentos, torna-se mais fácil interpretá-los, e gerar tipologias, que eventualmente podem se tornar populares. Por outro lado, em clusterização espacial, não necessariamente pretende-se chegar a um número pequeno de agrupamentos. O objetivo da clusterização espacial pode ser identificar municípios próximos e homogêneos para tornar, por exemplo, alguma política pública mais eficiente. Assim, é interessante ter *clusters* não muito grandes, para evitar longas distâncias entre municípios dentro do mesmo agrupamento. Portanto, interessa ao gestor público identificar muitos agrupamentos homogêneos com um número pequeno de unidades geográficas em cada um deles.

REFERÊNCIAS

ALPAYDIN, E. **Introduction to machine learning**. The MIT Press, 2004.

ANSELIN, **L. Spatial econometrics: methods and models.** Kluwer Academic, Dordrecth, 1988.

ANSELIN, L.; FLORAX, R. **Advances in spatial econometrics**. Heidelberg: Springer-Verlag, 2000.

ASSUNÇÃO, R.; LAGE, J.; REIS, E. Análise de conglomerados espaciais via árvore geradora mínima. **Revista Brasileira de Estatística,** vol. 63:220, p. 7-24, 2002.

BATAGELJ, V.; FERLIGOJ, A. **Constrained clustering problems.** Advances in Data Science and Classification, p. 137-144, 1998.

BERKHIN, P. **Survey of clustering data mining techniques**. Technical report, Accrue Software, San Jose, CA, 2002.

BERRY, M. J. A.; LINOFF, G. **Data mining techniques**. John Wiley and Sons, 1997.

CARVALHO, A. X. Y.; ALBUQUERQUE, C. W.; MOTA, J. A.; PIANCASTELLI, M. (Orgs.). **Dinâmica dos municípios**. Ipea, 2008.

CARVALHO, A. X. Y., ALMEIDA, G. R., ALBUQUERQUE, P. H. M., GUIMARÃES, R. D. **Clusterização hieráquica espacial.** Texto de Discussão do Ipea, Forthcoming, 2009.

CHEIN, F.; LEMOS, M. B.; ASSUNÇÃO, J. J. Desenvolvimento desigual: evidências para o Brasil. **Anais do Encontro Nacional de Economia**, 2005.

DUQUE, J. C.; RAMOS, R.; SURIÑACH, J. Supervised regionalization methods: a survey. **International Regional Science Review**, vol. 30, n. 3, p. 195-220, 2007.

ECK, J. E.; CHAINEY, S.; CAMERON, J. G.; LEITNER, M.; WILSON R. E. **Mapping crime**: understanding hot spots. U.S. Department of Justice, 2005.

GANGNON, R.; CLAYTON, M. K. **Cluster detection using Bayes factors from overparameterized cluster models**. Environmental and Ecological Statistics. 2007. Volume 14, Number 1 / March, 2007.

GLAZ, J., BALAKRISHNAN, N. **Scan statistics and applications**. Birkhäuser, 1999.

GLAZ, J.; NAUS, J.; WALLENSTEIN, S. **Scan statistics**. Springer, 2001.

GORDON, A. D. **A survey of constrained classification**. Computational Statistics and Data Analysis 21, p. 17-29, 1996

GOWER, J. C. A Comparison of Some Methods of Cluster Analysis. **Biometrics**, 1967.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The elements of statistical learning**: data mining, inference and prediction. Springer, 2001.

HIRSCHFIELD, A.; BOWERS, K. (Eds.). **Mapping and Analysing Crime Data**: lessons from research and practice. Taylor and Francis, 2001.

HIU, T.; D. FANG, J.; CHEN, Y.; WANG, JERIS C. A robust and scalable clustering algorithm for mixed type attributes in large database environment. *In*: **Proceedings of the seventh ACM SIGKDD international conference on knowledge discovery and data mining**. San Francisco, CA: ACM, 2001.

KHATTREE, R.; NAIK, D. N. **Multivariate data reduction and discrimination with SAS Software**. Wiley InterScience, 2000.

KULLDORFF, M. 1997. **A spatial scan statistic**. Communications in Statistics – Theory and methods 26: 1481-96.

LAWSON, A. B.; DENISON, D. G. T. (Eds.). **Spatial cluster modelling**. Chapman and Hall/CRC, 2002.

LI, X.; RAMACHANDRAN, R.; MOVVA, S.; GRAVES, S.; PLALE, B.; VIJAYAKUMAR, N. **Storm clustering for data-driven weather forecasting**. Technical report, University of Alabama, 2008.

LUO, Z. **Clustering under Spatial Contiguity Constraint**: a penalized K-means method. Technical Report, Department of Statistics, Penn State University, 2001.

MARAVALLE, M.; SIMEONE, B. **A spanning three heuristic for regional clustering**. Comm. Statist., Theory Methods, vol. 24, p. 629-63, 1995.

MARAVALLE, M.; SIMEONE, B.; NALDINI, R. **Clustering on Trees**. Computational Statistics and Data Analysis, vol. 24, p. 217-234, 1997.

PACE, K.; BARRY, R. **Sparse spatial autoregressions**. Statistics and Probability Letters, 33, 291-7, 1997.

SARLE, W. S. **Cubic clustering criterion**. SAS Technical Report A-108, Cary, NC: SAS Institute Inc, 1983.

WIPPERMAN, B. **Hierarchical Agglomerative Cluster Analysis with a contiguity constraint**. Simon Fraser University, 1999.

BIBLIOGRAFIA COMPLEMENTAR

CHOMITZ, K. M.; Da MATA, D.; CARVALHO, A.; MAGALHAES, J. C. R. **Spatial Dynamics of Labor Markets in Brazil**. World Bank Policy Research Working Paper 3752, 2005.

Da MATA, D.; DEICHMANN, HENDERSON, J. V.; LALL, S.; WANG, H. **Determinants of city growth in Brazil.** NBER Working Paper n. 11585, 2005a.

Da MATA, D.; PIN, C.; RESENDE, G. Composição e consolidação da infraestrutura domiciliar nos municípios brasileiros. Livro: Diferenças regionais no Brasil: caracterização e evolução nos últimos anos. Brasília: Ipea. (**no prelo**). 2006.

IBGE. Censo Demográfico 2000: documentação dos microdados da amostra. **Instituto Brasileiro de Geografia e Estatística**, 2002.

IPEA; PNUD; FJP. **Atlas do desenvolvimento humano no Brasil**. Brasília, 2003.

LUO, M.; MA, Y.; ZHANG, H. **A spatial constrained K-means approach to image segmentation**. Information, Communications and Signal Processing, 2003 and the Fourth Pacific Rim Conference on Multimedia. Vol. 2 p. 738 - 742, Dec., 2003.

MAHALANOBIS, P. C. **On the generalized distance in statistics**. *Proceedings of the National Institute of Sciences of India 2 (1): 49–55*, New Delhi, 1936.

MILLIGAN, G. W.; COOPER, M. C. An examination of procedures for determining the number of clusters in a Data Set. **Psychometrika**, 50,159 – 179, 1985.

MILLIGAN, G. W., COOPER, M.C. **A study of variable standardization**, College of Administrative Science Working Paper Series, 87 - 63, Columbus, OH: The Ohio State University, 1987.

MURTAGH, F.; A survey of algorithms for contiguity-constrained clustering and related problems. **The Computer Journal,** vol 28(1): p. 82-88, 1985.

APÊNDICE A1

RELAÇÃO DOS MUNICÍPIOS CRIADOS ENTRE 2000 E 2009

TABELA A1.1

Relação dos municípios criados entre 2000 e 2001

Código do município no ano de 2001	Nome do município no ano de 2001	População do município no ano de 2001
220779	Pau D'Arco do Piauí	3.031
240615	Jundiá	3.224
270375	Jequiá da Praia	12.848
290327	Barrocas	12.303
291955	Luís Eduardo Magalhães	19.213
320225	Governador Lindenberg	9.422
330285	Mesquita	168.042
430003	Aceguá	3.974
430047	Almirante Tamandaré do Sul	2.263
430107	Arroio do Padre	2.598
430222	Boa Vista do Cadeado	2.478
430223	Boa Vista do Incra	2.289
430258	Bozano	2.359
430461	Canudos do Vale	1.985
430462	Capão Bonito do Sul	1.919
430465	Capão do Cipó	2.566
430583	Coqueiro Baixo	1.594
430593	Coronel Pilar	1.921
430613	Cruzaltense	2.534
430843	Forquetinha	2.676
431065	Itati	2.878
431087	Jacuizinho	2.380
431123	Lagoa Bonita do Sul	2.467
431217	Mato Queimado	2.011
431346	Novo Xingu	1.835
431413	Paulo Bento	2.173
431417	Pedras Altas	2.571
431446	Pinhal da Serra	2.396
431453	Pinto Bandeira	2.636
431531	Quatro Irmãos	1.786
431595	Rolador	2.855
431673	Santa Cecília do Sul	1.724
431697	Santa Margarida do Sul	2.185
431861	São José do Sul	1.770
431936	São Pedro das Missões	1.794
432146	Tio Hugo	2.439
432377	Westfália	2.664
510185	Bom Jesus do Araguaia	3.880
510325	Colniza	10.909
510336	Conquista D'Oeste	2.639
510343	Curvelândia	4.584
510617	Nova Nazaré	1.982
510619	Nova Santa Helena	3.290
510631	Novo Santo Antônio	1.180
510757	Rondolândia	3.350
510774	Santa Cruz do Xingu	1.109
510776	Santa Rita do Trivelato	1.290
510779	Santo Antônio do Leste	1.930
510788	Serra Nova Dourada	1.064
510835	Vale de São Domingos	3.296
520485	Campo Limpo de Goiás	4.783
520815	Gameleira de Goiás	2.635
521015	Ipiranga de Goiás	2.808
521225	Lagoa Santa	923

Fonte: Datasus.
Elaboração dos autores.

TABELA A1.2

Relação dos municípios criados entre 2004 e 2005

Código do município no ano de 2005	Nome do município no ano de 2005	População do município no ano de 2005
220095	Aroeiras do Itaim	2.586
500390	Figueirão	2.930
510452	Ipiranga do Norte	2.530
510454	Itanhangá	4.195

Fonte: Datasus.
Elaboração dos autores.

APÊNDICE A2

LISTA DE DIVISÕES DA CLASSIFICAÇÃO NACIONAL DE ATIVIDADES ECONÔMICAS – CNAE 95

TABELA A2.1

Número total de estabelecimentos no Brasil por divisão CNAE 95

DIVISÃO CNAE 1995	TOTAL DE ESTABELECIMENTOS EM 2007
01 - Agricultura, pecuária e serviços relacionados	347.039
02 - Silvicultura, exploração florestal e serviços relacionados	14.670
05 - Pesca, aquicultura e serviços relacionados	4.946
10 - Extração de carvão mineral	649
11 - Extração de petróleo e serviços relacionados	894
13 - Extração de minerais metálicos	1.970
14 - Extração de minerais não metálicos	15.812
15 - Fabricação de produtos alimentícios e bebidas	97.126
16 - Fabricação de produtos do fumo	642
17 - Fabricação de produtos têxteis	22.340
18 - Confecção de artigos do vestuário e acessórios	94.565
19 - Preparação de couros e fabricação de artefatos de couro	26.457
20 - Fabricação de produtos de madeira	32.189
21 - Fabricação de celulose, papel e produtos de papel	7.252
22 - Edição, impressão e reprodução de gravações	46.973
23 - Fabricação de coque, refino de petróleo, elaboração de combustíveis	949
24 - Fabricação de produtos químicos	18.664
25 - Fabricação de artigos de borracha e plástico	22.535
26 - Fabricação de produtos de minerais não metálicos	35.917
27 - Metalurgia básica	8.353
28 - Fabricação de produtos de metal – exclusive máquinas e equipamentos	54.367
29 - Fabricação de maquinas e equipamentos	29.020
30 - Fabricação de maquinas para escritório e equipamentos de informática	1.261
31 - Fabricação de maquinas, aparelhos e materiais elétricos	9.378
32 - Fabricação de material eletrônico e de aparelhos e equipamentos de comunicações	2.600
33 - Fabricação de equipamentos de instrumentação para usos médico-hospitalares	5.747
34 - Fabricação e montagem de veículos automotores, reboques e carrocerias	7.157
35 - Fabricação de outros equipamentos de transporte	2.574
36 - Fabricação de móveis e indústrias diversas	47.166
37 – Reciclagem	3.783
40 - Eletricidade, gás e água quente	5.225
41 - Captação, tratamento e distribuição de água	3.243
45 – Construção	200.279
50 - Com. e rep. de veículos automotores e motocicletas, varejo de combustíveis	346.320
51 - Com. por atacado e representantes comerciais e agentes do comércio	341.637
52 - Com. varejista e reparação de objetos pessoais e domésticos	2.328.132
55 - Alojamento e alimentação	393.969
60 - Transporte terrestre	195.558
61 - Transporte aquaviário	2.148
62 - Transporte aéreo	1.935
63 - Atividades anexas e auxiliares do transporte e agências de viagem	61.990
64 - Correio e telecomunicações	27.378
65 - Intermediação financeira	49.809
66 - Seguros e previdência complementar	9.882
67 - Atividades auxiliares da intermediação financeira, seguros e previdência	38.473
70 - Atividades imobiliárias	222.449
71 - Aluguel de veículos, máquinas e equipamentos semicondutores	46.929
72 - Atividades de informática e serviços relacionados	133.686
73 - Pesquisa e desenvolvimento	2.483
74 - Serviços prestados principalmente às empresas	513.384
75 - Administração pública, defesa e seguridade social	21.632
80 – Educação	120.585
85 - Saúde e serviços sociais	230.501
90 - Limpeza urbana e esgoto e atividades relacionadas	5.072
91 - Atividades associativas	397.418
92 - Atividades recreativas, culturais e desportivas	117.785
93 - Serviços pessoais	94.881
95 - Serviços domésticos	11.251
99 - Organismos internacionais e outras instituições extraterritoriais	929
Total	6.887.958

Fonte: Ministério do Trabalho – RAIS.
Elaboração dos autores.

EDITORIAL

Coordenação
Iranilde Rego

Revisão
Cláudio Passos de Oliveira
Luciana Dias Jabbour
Marco Aurélio Dias Pires
Reginaldo da Silva Domingos
Leonardo Moreira de Souza (estagiário)
Maria Angela de Jesus Silva (estagiária)

Editoração
Bernar José Vieira
Cláudia Mattosinhos Cordeiro
Everson da Silva Moura
Renato Rodrigues Bueno

Livraria
SBS – Quadra 1 – Bloco J – Ed. BNDES, Térreo.
70076-900 – Brasília – DF
Fone: (61) 3315-5336
Correio eletrônico: livraria@ipea.gov.br
Tiragem: 130 exemplares