Carvalho, Alexandre Xavier Ywata; Albuquerque, Pedro Henrique Melo; de Almeida Junior, Gilberto Rezende; Guimarães, Rafael Dantas

Working Paper

Clusterização hierárquica espacial

Texto para Discussão, No. 1427

Provided in Cooperation with:
Institute of Applied Economic Research (ipea), Brasília

Suggested Citation: Carvalho, Alexandre Xavier Ywata; Albuquerque, Pedro Henrique Melo; de Almeida Junior, Gilberto Rezende; Guimarães, Rafael Dantas (2009) : Clusterização hierárquica espacial, Texto para Discussão, No. 1427, Instituto de Pesquisa Econômica Aplicada (IPEA), Brasília

This Version is available at:
https://hdl.handle.net/10419/91235

TEXTO PARA DISCUSSÃO Nº 1427

CLUSTERIZAÇÃO HIERÁRQUICA ESPACIAL

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

TEXTO PARA DISCUSSÃO Nº 1427

CLUSTERIZAÇÃO HIERÁRQUICA ESPACIAL*

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

Brasília, outubro de 2009

* Os autores agradecem as sugestões de Bruno de Oliveira Cruz e Paulo Furtado de Castro. 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 (endereço eletrônico: alexandre.ywata@ipea.gov.br).
*** 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 estudo apresenta uma nova metodologia para clusterização hierárquica espacial de polígonos contíguos, com base em um sistema de coordenadas georreferenciadas. O algoritmo proposto é construído a partir de uma modificação do algoritmo de clusterização hierárquica tradicional, comumente utilizado na literatura de análise multivariada. De acordo com o método proposto neste trabalho, 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 como vizinhos polígonos que possuem um vértice em comum (vizinhança do tipo *queen*) ou uma aresta em comum (vizinhança do tipo *rook*). O estudo apresenta aplicações da nova metodologia para clusterização dos municípios brasileiros, no ano de 2000, com base em um conjunto de variáveis socioeconômicas. Diversos métodos de clusterização são estudados, assim como diferentes tipos de distâncias entre vetores. Os métodos estudados foram: *centroid*, *single linkage*, *complete linkage*, *average linkage* e *average linkage weighted*, *Ward's minimum variance* e método da mediana. As distâncias utilizadas foram: norma L_p (em particular, as normas L_1 e L_2), Mahalanobis e distância euclidiana corrigida pela variância (*variance corrected*) – caso particular da distância de Mahalanobis. Finalmente, apresenta-se uma discussão sobre alguns métodos comumente utilizados para seleção do número de *clusters*.

ABSTRACT

This paper presents a new methodology for hierarchical spatial clustering of contiguous polygons, based on a geographic coordinate system. The proposed algorithm is built upon a modification of traditional 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, for the year 2000, based on a group of socio-economic variables. Several clustering methods are investigated, as well as several types of vector distances. 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: L_p norm (particularly, L_1 e L_2 norms), Mahalanobis distance and variance corrected Euclidian distance. Finally, a discussion on selection of the number of clusters is presented.

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 ou a diversidades climáticas e geográficas. Estas diferenças espaciais podem surgir devido a grandes dimensões continentais (no caso de alguns países) ou a particularidades no histórico político e econômico de cada região. A 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.

Nas políticas de desenvolvimento intraurbano, mesmo lidando-se com dimensões bem menores quando comparadas às dimensões do território nacional, o problema da heterogeneidade 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 do tratamento adequado da heterogeneidade espacial, do ponto de vista das 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 desmatamento tenha sido mais intenso nas últimas décadas. Estudos de epidemiologia exigem que o pesquisador possa detectar, com certa precisão, áreas de maior influência de uma determinada 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 técnicas de tratamento de dados mais populares e eficientes na 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 de livros de estatística multivariada. A ideia de agrupar dados que compartilham 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 Hastie, Tibshirani e Friedman (2001), Khattree e Naik (2000), Berry e Linoff (1997), e Alpaydin (2004).

Especificamente no que respeita a dados geográficos, a clusterização espacial é uma técnica poderosa, adaptável a diversos casos, razão pela qual 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 agrupamento espacial para aprimorar a previsão do tempo. Na área de saúde, Gangnon e Clayton (2003) estudaram casos de leucemia em Nova York para aprimorar métodos de clusterização espacial, usando 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, número de casos de leucemia por habitante etc.).

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) de onde os crimes aconteceram, e são empregadas técnicas de estimação de densidade não paramétrica, em duas dimensões; os *hot spots* correspondem a áreas no espaço onde 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 – reside apenas em identificar locais nos quais 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; GLAZ *et al.*, 2001). Ao invés de ser utilizado para o tratamento 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 extrapole certo limite, comparando-se o resultado com alguma distribuição apropriada para o tipo de dados em questão (*e.g.* Bernoulli para dados binários). Trata-se de uma técnica computacionalmente intensiva, que tem sido bastante utilizada para detecção de epidemias.

Neste estudo, apresenta-se uma metodologia para análise de dados espaciais que é conceitualmente diferente das técnicas *hot spots* e *scan*. Ao invés de tentar detectar áreas onde a ocorrência de determinado evento seja significantemente mais pronunciada, a técnica abordada neste trabalho 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 georreferenciado.

Pode-se estar interessado em analisar a heterogeneidade espacial dos municípios brasileiros (polígonos) de acordo com a renda *per capita*, longevidade, escolaridade média e condições dos domicílios (variáveis características). Denominamos o método apresentado de *clusterização hierárquica espacial*, o qual corresponde a uma modificação dos algoritmos de clusterização hierárquica tradicionais (HASTIE, TIBSHIRANI e FRIEDMAN, 2001; KHATTREE e NAIK, 2000; BERRY e LINOFF, 1997).

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 *cluster* e atenta para suas especificidades. A clusterização hierárquica pode ser feita de forma aglomerativa (iniciando com tantos *clusters* quantos forem os objetos, e então os unindo 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, que relaciona as distâncias de cada objeto (vetor de dados) a cada outro objeto. No caso da clusterização hierárquica aglomerativa, os objetos próximos são unidos em *clusters* e a matriz de distâncias é atualizada. O processo interage até o número estabelecido de *clusters*.

De acordo com os algoritmos de clusterização hierárquica espacial propostos neste estudo, 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 que também sejam importantes para a caracterização das unidades geográficas não estejam incluídas na base, o que incorreria 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 BARRY, 1997). Nesse caso, a utilização de algoritmos de clusterização quando a contiguidade é imposta pode reduzir a perda de informação devida 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 justamente identificar regiões homogêneas no país ou dentro de uma área urbana, nas quais políticas de desenvolvimento diferenciadas possam ser implementadas. Dessa forma, a

contiguidade é fundamental, pois a intenção é a formulação de políticas públicas focadas em áreas geográficas que apresentem algum grau de vizinhança.

Os três desafios da clusterização são estabelecer a medida de similaridade ou dissimilaridade empregada, a distância entre vetores de dados, e definir o número final de *clusters*. As medidas de dissimilaridade empregadas neste estudo são: *average linkage*, *centroid*, *single linkage*, *complete linkage (unweighted)*, *complete linkage (weighted)*, *Ward's minimum variance* e método da mediana. Além das diferentes medidas de dissimilaridade, empregaram-se também diferentes distâncias entre vetores: 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*). Por fim, traremos a discussão a respeito da definição do número de *clusters*, utilizando os critérios *cubic clustering criterion (CCC)*, pseudo-*F*, pseudo-t^2, R^2 e R^2 semiparcial.

Algoritmos alternativos para a 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. 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 estudo, os algoritmos hierárquicos 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. Por outro lado, a nova metodologia permite a incorporação imediata de diferentes medidas de dissimilaridade entre grupos homogêneos. Além disso, critérios tradicionais de escolha do número (estatísticas *CCC*, pseudo-*F* e pseudo-t^2, R^2 e R^2 semi-parcial) de *clusters* também podem ser facilmente incorporados.

O trabalho está dividido em cinco seções, incluindo esta introdução. A segunda parte aborda a metodologia utilizada para a formação dos grupos homogêneos de municípios (*clusters*), a discussão sobre as medidas de dissimilaridade, o cálculo das diferentes distâncias, e a 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 dados socioeconômicos para os municípios brasileiros, com referência na malha de municípios de 2000. A quarta seção é reservada às conclusões do trabalho.

2 METODOLOGIA

Nesta seção, o algoritmo para formação dos agrupamentos homogêneos de municípios é descrito. Conforme será abordado em mais detalhes a seguir, o algoritmo utilizado neste trabalho corresponde a uma modificação dos algoritmos de clusterização hierárquica comumente apresentados na literatura

2.1 ALGORITMO PARA A 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 que o analista ou pesquisador esteja 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. Depois, discutem-se as modificações no método de clusterização tradicional, de forma a incorporar a restrição de unidades geográficas (por exemplo, municípios, setores censitários, Unidades da Federação, áreas de ponderação) contíguas.

2.2 ALGORITMOS DE CLUSTERIZAÇÃO HIERÁRQUICA

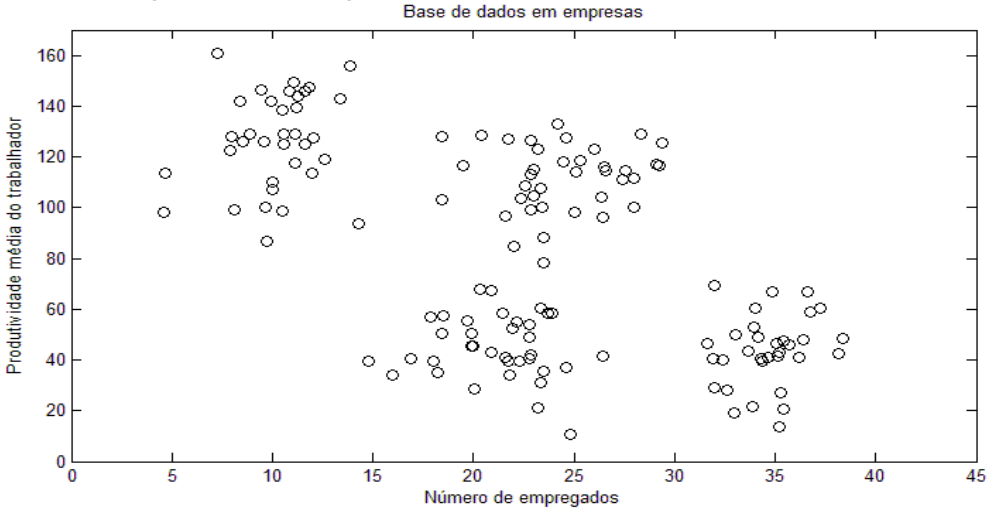
Para exemplificar a ideia geral dos algoritmos de clusterização, considere a figura 2.1, contendo 141 observações, cada qual correspondendo a uma empresa específica (base de dados hipotética, para fins de ilustração). O eixo horizontal do gráfico indica o número de empregados em cada uma das empresas, 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 que o grupo 1, em vermelho,[1] pode ser interpretado como sendo o grupo de empresas com baixo número de empregados e alta produtividade. O grupo 2, em azul, pode ser considerado o grupo com baixo número de empregados e baixa produtividade. O grupo 3, em preto, corresponde às empresas de médio porte e 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>.

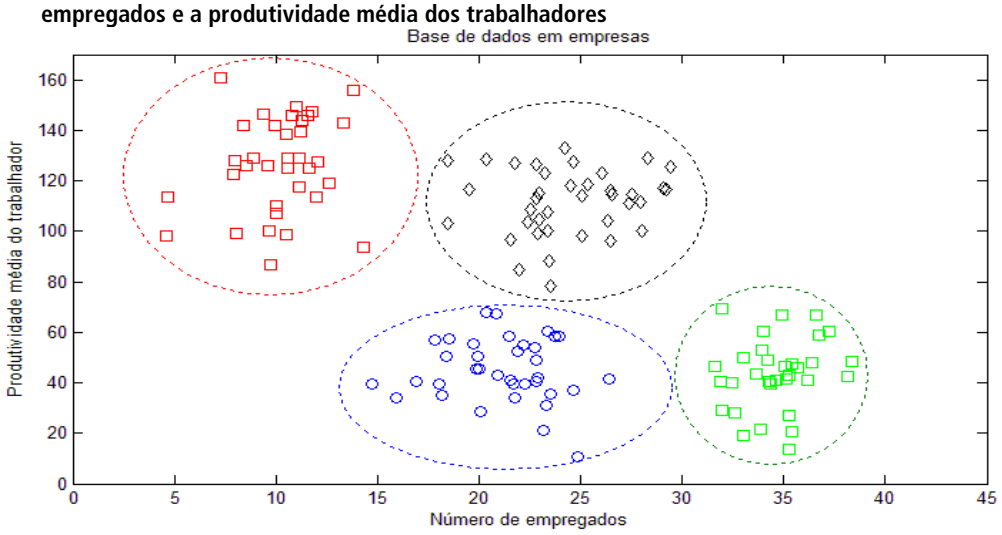
FIGURA 2.1
Informações sobre número de empregados e produtividade média do trabalhador para uma base de dados hipotética de 141 empresas



Elaboração dos autores.

Dada a grande importância do problema da 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)* modelos de mistura (*mixture models*); e *iii)* busca por modas (*mode seeking*). As últimas duas categorias baseiam-se em alguma forma de modelo probabilístico para o processo gerador de dados. Já os algoritmos combinatórios podem ser vistos basicamente como regras heurísticas de busca dos melhores agrupamentos de observações. De forma geral, não existe um algoritmo que seja superior aos demais em todas as situações. A decisão acerca de qual deles melhor se aplicará ao caso em questão dependerá do processo gerador de dados, bem como da experiência do analista ou pesquisador e da disponibilidade de *softwares* específicos.

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.

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

1. Considere-se uma base inicial de *N clusters*. Em geral, estes agrupamentos correspondem simplesmente às unidades a serem agrupadas em grupos 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 $x_i = [x_{i,1}\ x_{i,2}\ ...\ x_{i,m}]$. Estas características podem ser características socioeconômicas, por exemplo.

2. Calcula-se a distância entre todos os pares formados por elementos dentre esses *N clusters* iniciais. *Distância*, neste caso, pode ser qualquer medida de dissimilaridade entre o conjunto de atributos $x_i = [x_{i,1}\ x_{i,2}\ ...\ x_{i,m}]$. 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, podemos citar a medida de Ward, que pode ser escrita como:

$$D_{K,L} = \frac{\left\| \overline{X}_K - \overline{X}_L \right\|^2}{\left(\frac{1}{N_K} + \frac{1}{N_L} \right)}$$

   onde $D_{K,L}$ é a medida de dissimilaridade (ou distância) entre o *cluster L* e o *cluster K*, $\overline{X_L}$ e $\overline{X_K}$ são os vetores correspondentes às médias dos vetores de características de todas as unidades (municípios, por exemplo) dentro dos *clusters L* e *K*, respectivamente, 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 função corresponde à norma euclidiana. A seção 2.4 apresenta um conjunto de outras medidas de dissimilaridade (além de uma discussão mais detalhada sobre a medida de dissimilaridade de Ward), bem como um conjunto de distâncias entre vetores (incluindo a norma euclidiana).

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, como 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. Isto reduz enormemente o tempo de processamento.

O passo final é 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 a seleção do número de agrupamentos foram propostas, sem ter havido necessariamente um consenso sobre qual medida utilizar. Algumas destas 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 supracitados, indicando portanto quantos *clusters* utilizar. A seção 2.5 traz 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-se uma interpretação plausível para todos os *clusters* formados. Isto 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, entre outras, sequências de DNA. Neste trabalho, as unidades a serem agrupadas são municípios ou setores censitários, 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 a 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 trabalho, estas subunidades correspondem a municípios). Inicialmente, cada uma dessas *N* observações consiste isoladamente em um *cluster*, 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, investigaram-se 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 conhecido, 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 conhecido como vizinhança do tipo *queen*. Obviamente, a vizinhança do tipo *queen* é menos restritiva 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 $N \times (N-1)/2$, 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 $[x_{i,1}\ x_{i,2}\ \ldots\ x_{i,m}]$, 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 $[x_{i,1}\ x_{i,2} \ldots\ x_{i,m}]$. Em geral, estas variáveis correspondem a características socioeconômicas – como é o 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 ocorre 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 para se analisar o comportamento dos indicadores. Por outro lado, a utilização direta de critérios estatísticos não necessariamente implicará um número de *clusters* que faça sentido em relação aos objetivos de cada projeto. Pode-se optar por selecionar o número de agrupamentos cuja interpretação econômica faça mais sentido para os objetivos do trabalho. A escolha do número de agrupamentos homogêneos via critérios subjetivos foi utilizada, por exemplo, em Chein, Lemos e Assunção (2005), no qual foram selecionados 100 *clusters* para todo o território brasileiro. Em todo caso, a seção 3.3 apresenta uma discussão sobre o comportamento de alguns dos critérios de seleção do número de *clusters*, comumente encontrados na literatura.

2.4 MEDIDAS DE DISSIMILARIDADE ENTRE *CLUSTERS*

Nesta seção, serão apresentadas 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 não é exaustiva, podendo o leitor recorrer às referências apresentadas neste estudo 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. As distâncias podem ser combinadas com as medidas de dissimilaridade para gerar uma grande variedade de métodos possíveis. Estas várias combinações serão estudadas na seção 3.2.

2.4.1 MEDIDAS DE DISSIMILARIDADE

As medidas de dissimilaridade estudadas neste trabalho estão listadas a seguir. Além da apresentação 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 como método de McQuitty, define a distância média entre pares de objetos como sendo a relevante para a elaboração da matriz de distâncias. É um método que tende a juntar *clusters* com baixa variância, sendo 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 ocorre por intermédio da atualização da matriz de dissimilaridade entre *clusters*. Toda vez que um novo *cluster* C_M é criado, a matriz de dissimilaridades é atualizada a partir da junção dos *clusters* C_L e C_K existentes no passo anterior, para considerar as distâncias ao novo *cluster*. Esta atualização pode ser feita diretamente com base nas distâncias existentes na matriz no passo anterior, utilizando-se fórmulas combinatórias (*combinatorial formulas*). Considere 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. Como consequência, ele é menos influenciado por valores extremos, como é o caso dos métodos *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* que estejam mais próximos entre si. Possui muitas propriedades teóricas desejáveis, mas tem desempenho ruim 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:

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

Uma vez que ele não impõe restrições à forma dos *clusters*, este método sacrifica a possibilidade de se obterem *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* que estejam 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* estejam 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 no sentido 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 *clusters* unidos, detectando qual união produz o menor aumento de *SQE*. Neste método, a distância entre dois *clusters* é dada pela soma de quadrados Anova (*analysis of variance)* entre os dois *clusters*, para todas as variáveis. A cada geração, minimiza-se a soma de quadrados *intra-cluster* obtenível pela união de dois *clusters*. Frequentemente, aconselha-se utilizar a razão *SQE*/*SQET* no lugar do *SQE* absoluto. A medida de dissimilaridade é definida por:

$$D_{K,L} = \frac{d(\bar{x}_K, \bar{x}_L)^2}{\left(\frac{1}{N_K} + \frac{1}{N_L} \right)}$$

No caso de utilizar-se a distância euclidiana, a equação acima pode ser reescrita como (equação tradicionalmente encontrada na literatura de clusterização):

$$D_{K,L} = \frac{\left\| \bar{x}_K - \bar{x}_L \right\|^2}{\left(\frac{1}{N_K} + \frac{1}{N_L} \right)}$$

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, sendo fortemente viesado no sentido 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* como sendo o quadrado da distância euclidiana entre os centroides dos *clusters*. A distância entre os *clusters* é definida por:

$$D_{K,L} = d(\bar{x}_K, \bar{x}_L)^2$$

Como se trata de uma comparação de médias, *outliers* exercem pouca influência. Em outros aspectos, pode perder em eficiência para os métodos *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:

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

onde 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

Apresentarão-se a seguir os tipos de distâncias topológicas que serão utilizadas neste estudo. 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 de 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 renda *per capita*, longevidade média, escolaridade média e índice de Gini do município. Utilizaremos a notação x_i para especificar o i-ésimo elemento (escalar) do vetor x.

Norma L_p

A Norma L_p é o tipo de métrica mais usualmente utilizada, na qual o parâmetro p é um fator de penalização para dados extremos (*outliers*). De acordo com a norma L_p, a distância entre os vetores x e y é definida como:

$$d(x, y) = \sqrt[p]{\sum_{i=1}^{n} \left| x_i - y_i \right|^p}$$

Distância euclidiana (Norma L_2)

A norma ou distância euclidiana é um caso particular da nova L_p para $p = 2$. A distância euclidiana entre os vetores x e y é definida como:

$$d(x,y) = \sqrt[2]{\sum_{i=1}^{n} \left| x_i - y_i \right|^2}$$

Norma L1(Manhattan, taxicab ou city block distance)

A norma ou distância euclidiana é um caso particular da nova L_p para $p = 1$. A distância euclidiana entre os vetores x e y é definida como:

$$d(x,y) = \sum_{i=1}^{n} \left| x_i - y_i \right|$$

Distância de Mahalanobis

A distância de Mahalanobis pode ser vista como uma generalização da forma quadrática da distância euclidiana (MAHALANOBIS, 1936). Difere da distância euclidiana por ser invariante à escala, isto é, não depender da escala das medidas em estudo. Seja $\sum$ a matriz de variância-covariância para as variáveis que caracterizam os polígonos a serem agrupados. Para vetores de dados com v variáveis, a matriz $\sum$ é uma matriz simétrica, com dimensão $v \times v$. Nas aplicações, a matriz $\sum$ é calculada diretamente a partir da base de dados. Uma vez calculada a matriz $\sum$, a distância de Mahalanobis entre os vetores x e y é definida como:

$$d(x,y) = \sqrt{(x-y)' \sum^{-1} (x-y)}$$

No caso particular no qual a matriz de variância-covariância é a matriz identidade, a distância de Mahalanobis coincide com a distância euclidiana.

Distância euclidiana corrigida pela variância (variance corrected) ou distância euclidiana normalizada

A distância corrigida pela variância corresponde a um caso particular da distância de Mahalanobis, para o caso no qual a matriz $\sum$ é uma matriz diagonal (elementos nulos fora da diagonal principal) e os elementos da diagonal principal são as variâncias das v variáveis do vetor de dados, caracterizando cada polígono. Neste caso, a utilização da distância *variance corrected* corresponde à utilização da distância euclidiana, na qual todas as variáveis da base original são divididas pelos seus respectivos desvios padrões. Uma das vantagens de se utilizar a distância corrigida pela variância é a correção de problemas de escalas entre as variáveis. A depender da medida de dissimilaridade utilizada, variáveis com escala maior podem acabar tendo artificialmente um peso maior na formação dos *clusters*. Por outro lado, conforme discutido em Hastie, Tibshirani e Friedman (2001), não necessariamente colocar todas as variáveis na mesma escala, por meio da divisão pelos respectivos desvios padrões, traz benefícios na identificação dos grupos homogêneos. Vide Milligan e Cooper (1987) para uma discussão sobre normalização das variáveis para clusterização.

2.4.3 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* que serão 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 para o *CCC* produzem a rejeição de H_0. Plotam-se os valores do *CCC* e o número de *clusters* e procuram-se por picos nos quais *CCC* excede três. 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 esperado 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 por valores de pico na sequência de estatísticas pseudo-I 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 como um só, enquanto valores altos para o critério R^2 semiparcial indicam que os *clusters* unidos são provavelmente diferentes. Para mais detalhes sobre os diversos critérios de escolha do número de *clusters*, vide Khattree e Naik (2000).

3 ESTUDO DE CASO

Esta seção apresenta um estudo de caso para analisarem-se as propriedades das medidas de dissimilaridade apresentadas na seção 2.4. A base de dados utilizada refere-se a uma base de informações socioeconômicas municipais do Brasil, com base na malha de municípios do ano de 2000. Ao todo, são 5.507 municípios. A escolha desta base foi motivada pela necessidade, no Brasil, de estudos e políticas públicas focadas em áreas homogêneas e contíguas do território nacional. Portanto, o estudo de caso serve não somente para apresentar algumas indicações gerais sobre as diversas medidas de dissimilaridade estudadas, mas também para balizar qual medida de

dissimilaridade é mais aconselhável para estudos de desenvolvimento regional específicos. Na seção 3.1, discute-se a base de dados socioeconômicos utilizada. A seção 3.2 discute os resultados do exercício de clusterização espacial, comparando diferentes medidas de dissimilaridade. Os resultados sobre a investigação do comportamento dos critérios de seleção do número de *clusters* são apresentados na seção 3.3.

3.1 BASE DE DADOS

Os dados utilizados neste trabalho são oriundos do *Censo Demográfico 2000* (IBGE, 2002) e do *Atlas do Desenvolvimento Humano no Brasil* (IPEA, PNUD e FJP, 2003). Para uma discussão mais aprofundada sobre os dados utilizados, vide Carvalho, Albuquerque, Mota e Piancastelli (2008). Diversas características socioeconômicas dos municípios brasileiros foram selecionadas, a saber:

a) Taxa de emprego do município (população empregada divida pela população total).

b) Percentual da população do município em áreas urbanas.

c) Variáveis demográficas: longevidade e taxa de fecundidade do município em 2000.

d) Infraestrutura urbana e condições dos domicílios: percentual de domicílios com iluminação pública, identificação (CEP), esgotamento sanitário, água encanada, pavimentação, energia elétrica e coleta de lixo.

e) Desempenho educacional: percentual de crianças de 5 a 6 anos na escola; percentual de crianças de 7 a 14 anos com acesso ao curso fundamental; percentual de adolescentes de 15 a 17 anos com acesso ao ensino médio; percentual de pessoas de 18 a 24 anos com acesso ao curso superior; percentual de crianças de 7 a 14 anos com mais de um ano de atraso escolar; percentual de professores do ensino fundamental residentes com curso superior; média de anos de estudo das pessoas de 25 anos ou mais de idade; percentual de pessoas de 15 anos ou mais analfabetas; percentual de crianças de 10 a 14 anos analfabetas.

f) Renda domiciliar *per capita*, percentual da renda domiciliar *per capita* proveniente de rendimentos do trabalho e percentual da renda proveniente de transferências governamentais em 2000.

g) Percentual de pobres no município em 2000.

h) Desigualdade de renda do domicílio (índice de Gini) em 2000.

i) Variáveis relacionadas à saúde pública do município: mortalidade infantil de crianças até 1 ano e de crianças até 5 anos em 2000, e probabilidade de sobrevivência até os 60 anos em 2000.

j) Taxa de homicídios no município em 2002.

Além dos dados descritos, foi utilizada a malha de municípios brasileiros de 2000, contendo informações georreferenciadas. Estas informações foram utilizadas para construir a estrutura de vizinhança entre os municípios. Devido à presença de ilhas no território nacional, o processo de agregação sequencial de *clusters* prosseguiu até atingir-se 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*).

3.2 EFEITOS DO TIPO DE DISSIMILARIDADE E DO TIPO DE DISTÂNCIA ENTRE VETORES

O anexo 1 apresenta os mapas dos agrupamentos espaciais para as diferentes medidas de dissimilaridade, as diferentes distâncias e os diferentes tipos de vizinhança. 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. Para complementar a análise, o anexo 2 apresenta os *box plots* pra avaliar o tamanho dos *clusters* formados.

Observando-se os mapas, nota-se que os métodos *single linkage*, *average linkage*, *average linkage weighted*, *centroid* e da mediana formam *clusters* bastante diferentes em termos de número de unidades geográficas. Esta conclusão é corroborada pelos *box plots* no anexo 2. O método *complete linkage*, e principalmente o método de Ward, tendem a formar *clusters* com tamanhos mais homogêneos. Por outro lado, entre as cinco distâncias estudadas, a distância Manhattan (L_1) é a que fornece *clusters* de tamanhos mais homogêneos. Nota-se que, para a distância geral L_p, quando mais próximo de $p = 1$, mais homogêneo será o tamanho dos agrupamentos. Isto pode estar relacionado ao fato de a distância L_1 atribuir menos peso às observações discrepantes do que a distância euclidiana, por exemplo.

3.3 COMPORTAMENTO DOS CRITÉRIOS DE SELEÇÃO DO NÚMERO DE *CLUSTERS*

O anexo 3 apresenta os gráficos dos critérios de seleção para o número de agrupamentos homogêneos. Os gráficos referem-se aos sete métodos, considerando-se apenas a distância euclidiana. Para as demais distâncias, as conclusões são similares. Os critérios apresentados nos gráficos são o R^2, o *CCC*, a estatística pseudo-t^2 e a estatística pseudo-*F*. O R^2 semiparcial apresentou comportamento semelhante ao critério pseudo-t^2 e não foi apresentado.

No gráfico para a estatística R^2, apresenta-se também o R^2 esperado (*expected* R^2). Este valor é o mesmo para todos os métodos de clusterização. Nota-se que o *expected* R^2 está bem acima dos R^2 para as sete medidas de dissimilaridade. Isto já era esperado: o *expected* R^2 é calculado (SARLE, 1983) sob a condição de ausência de restrição de contiguidade entre os *clusters*, enquanto o R^2 para nos algoritmos de clusterização espacial tendem a ser menores do que o R^2 nos algoritmos de clusterização tradicionais. Este fato faz com que os valores para o critério *CCC* sejam bastante negativos (vide expressão para o *CCC*); portanto, a regra de se escolher um número de *clusters* no qual o *CCC* apresente picos acima do valor três não é mais válida. Isto suscita uma pergunta para pesquisas futuras: como calcular o R^2 esperado de forma mais apropriada para o caso de clusterização espacial hierárquica.

Os critérios pseudo-t^2 e pseudo-*F* apresentam um comportamento gráfico similar ao caso dos algoritmos hierárquicos não espaciais. Note-se que o critério pseudo-*F* apresenta uma sequência com variações não muito abruptas, contendo alguns máximos locais, que podem ser indicativos do número de *clusters* a ser escolhido. Por outro lado, a sequência para o critério pseudo-t^2 apresenta alguns pontos de grandes picos. Por este motivo, utilizou-se a escala logarítmica para o eixo vertical nos gráficos

do critério pseudo-t^2. A literatura sugere que a escolha do número de *clusters* seja igual a *u+1*, onde *u* é o número de *clusters* para o qual existe um pico. Portanto, o critério pseudo-t^2 parece apresentar evidências mais claras para a escolha do número de *clusters* a serem utilizados.

3.4 EFEITOS DO TIPO DE VIZINHANÇA (*ROOK VERSUS QUEEN*)

Observando-se os mapas no anexo 2, nota-se que, em se tratando da distribuição do número de municípios em cada *cluster*, os resultados utilizando-se a vizinhança do tipo *rook* são bem semelhantes aos resultados utilizando-se a vizinhança do tipo *queen*. A mesma semelhança vale para os critérios de seleção do número de *clusters*, apresentados nos gráficos do anexo 3. Por outro lado, o anexo 1 mostra que os agrupamentos formados com as duas distâncias podem ser bastante diferentes, mesmo 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 a vizinhança, pode implicar a 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 *intra-clusters*. No exercício apresentado na próxima seção, no entanto, esta menor variabilidade para a vizinhança do tipo *queen* não foi verificada.

3.5 COMPARAÇÃO COM AGRUPAMENTOS POLÍTICOS DE MUNICÍPIOS BRASILEIROS

Nesta seção, apresenta-se uma comparação dos agrupamentos obtidos via clusterização espacial hierárquica com agrupamentos políticos de municípios existentes no Brasil. Os agrupamentos utilizados, para fins de comparação, são: microrregiões, mesorregiões e Unidades da Federação. Ao todo, são 27 Unidades da Federação, 558 microrregiões e 137 mesorregiões. Portanto, para compararem-se 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 como indicador de *performance* da configuração de agrupamentos – ela fornece uma ideia da variabilidade *intra-clusters* 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 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 3.1 apresenta a medida de *performance* para diferentes tipos de medidas de dissimilaridade e diferentes distâncias entre vetores. Note que os *clusters* obtidos via clusterização espacial hierárquica, para o método de Ward, apresentam menores variabilidades totais do que os agrupamentos políticos. Esta observação vale tanto para microrregiões (558 *clusters*) como para 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 50%. Para o método *complete linkage*, a medida de variância *intra-clusters* também é menor, em geral, do que a medida de variabilidade *intra-clusters* utilizando-se a divisão política; a exceção ocorre quando se utiliza a distância de Mahalanobis. Para os demais cinco métodos, os resultados são pouco encorajadores; a variabilidade resultante com as medidas de dissimilaridade *single linkage*, *average linkage*, *average linkage weighted*, da mediana e *centroid* são maiores do que a variabilidade dada pelas divisões políticas, no caso de Unidades da Federação e mesorregiões. Para microrregiões, no entanto, os métodos *average linkage* e *average linkage weigthed* apresentam, para a maior das distâncias, variabilidades menores que as divisões políticas.

O fato de os métodos *complete linkage*, e principalmente o método de Ward, terem gerados *clusters* com menor variabilidade que as divisões políticas não significa que os *clusters* obtidos via algoritmos numéricos sejam 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 que sejam 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.

Para a maioria dos métodos, os algoritmos de clusterização espacial não incorreram em agrupamentos mais homogêneos, do ponto de vista da medida de variabilidade *intra-clusters* utilizada, que os agrupamentos políticos. Isto pode ser explicado pelo fato de estes métodos tenderem à formação de agrupamentos muito desiguais em número de municípios (seção 3.2). Por isso, alguns dos *clusters* contêm mais da metade dos municípios brasileiros, para os quais a variabilidade agregada é muito alta. No agregado, a variabilidade total para todos os agrupamentos acaba sendo bem mais alta que a variabilidade para os agrupamentos políticos (microrregião, mesorregião ou Unidades da Federação).

TABELA 3.1

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	27		137		558	
		rook	*queen*	*rook*	*queen*	*rook*	*queen*
Single linkage	Euclidiana	214	214	259	259	297	297
	L_1 - Manhattan	216	216	265	265	305	305
	L_p (p = 1,5)	215	215	263	263	302	302
	Euclidiana normalizada	217	217	265	265	301	301
	Mahalanobis	216	216	265	265	296	296
Complete linkage	Euclidiana	91	96	77	80	56	56
	L_1 - Manhattan	93	91	77	77	59	58
	L_p (p = 1,5)	97	92	74	77	56	56
	Euclidiana normalizada	101	97	84	83	66	66
	Mahalanobis	156	187	100	97	85	84
Average linkage (unweighted)	Euclidiana	215	218	133	148	97	106
	L_1 - Manhattan	195	215	175	127	88	84
	L_p (p = 1,5)	189	127	136	118	90	84
	Euclidiana normalizada	118	147	137	139	100	96
	Mahalanobis	164	137	117	116	98	98
Average linkage (weighted)	Euclidiana	204	203	110	106	71	71
	L_1 - Manhattan	110	112	112	115	73	75
	L_p (p = 1,5)	205	205	114	114	74	71
	Euclidiana normalizada	209	209	125	125	93	91
	Mahalanobis	212	212	251	249	150	153
Mediana	Euclidiana	199	219	244	259	223	150
	L_1 - Manhattan	219	220	251	154	201	158
	L_p (p = 1,5)	218	198	175	240	183	255
	Euclidiana normalizada	217	217	193	170	191	176
	Mahalanobis	218	215	160	162	161	173
Centroid	Euclidiana	204	205	119	109	101	96
	L_1 - Manhattan	212	212	252	251	132	132
	L_p (p = 1,5)	204	204	121	121	118	118
	Euclidiana normalizada	211	211	133	134	132	133
	Mahalanobis	212	212	257	257	283	284
Ward	Euclidiana	77	76	66	66	53	52
	L_1 - Manhattan	80	79	72	71	57	56
	L_p (p = 1,5)	78	78	66	66	53	53
	Euclidiana normalizada	85	85	76	76	63	63
	Mahalanobis	90	89	86	86	77	76
Divisão política de municípios	Unidades da Federação	100		---		---	
	Mesorregiões	---		100		---	
	Microrregiões	---		---		100	

Elaboração dos autores.

O método de clusterização hierárquica que resultou na menor medida de variabilidade foi o método de Ward, utilizando a distância euclidiana. Isto já era esperado, pois, conforme relatado na literatura de clusterização hierárquica tradicional (não hierárquica), o algoritmo de Ward tende a formar *clusters* de forma que a variância agregada seja minimizada. Além disso, a medida de variabilidade (*TCSS*) utilizada foi construída utilizando-se intrinsecamente a norma euclidiana. Lançando-se mão de outras medidas de variabilidade, mais ligadas a outros tipos de distância, a variabilidade resultante possivelmente seria menor no caso de métodos de clusterização baseados na distância L_1, por exemplo, ou na distância de Mahalanobis. Em todo caso, este exercício reforça a utilização dos algoritmos de clusterização hierárquica espacial, sempre que possível, com preferência para o método de Ward.

Vale notar que, mesmo permitindo maior flexibilização na formação dos *clusters*, a vizinhança do tipo *queen* não necessariamente implica agrupamentos com menor variabilidade que os agrupamentos obtidos com a vizinhança do tipo *rook*. Isto pode

ser explicado pelo caráter hierárquico dos algoritmos de clusterização utilizados neste estudo. Se estivéssemos utilizando um algoritmo de minimização da variabilidade total, possivelmente a vizinhança do tipo *queen* incorreria em uma variabilidade menor 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.

4 CONCLUSÕES

Este estudo apresentou uma nova metodologia para a formação hierárquica de agrupamentos espaciais. O algoritmo proposto diz respeito a uma modificação do algoritmo de clusterização hierárquica tradicional: a cada passo do processo de junção de dois *clusters*, impõem-se, para formação de um novo, que a junção possa acontecer somente entre *clusters* geograficamente vizinhos (de acordo com um sistema de dados geoferrenciados). Neste caso, consideraram-se dois tipos de vizinhanças: do tipo *rook* (polígonos com uma aresta em comum); e do tipo *queen* (polígonos com um vértice em comum). Devido ao fato de o algoritmo de clusterização especial hierárquica proposto neste estudo 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 estudo foram: *Ward's minimum variance*, *centroid*, mediana, *single linkage*, *complete linkage*, *average linkage*, e *average linkage weigthed*. Além disso, empregaram-se diferentes definições de distâncias entre vetores. As distâncias empregadas foram: distância L_p, distância euclidiana, distância L_1, distância de Mahalanobis e distância euclidiana corrigida pela variância (*variance corrected*).

A partir de um estudo de caso detalhado, que utilizou variáveis socioeconômicas para os municípios brasileiros no ano de 2000, diversas propriedades dos métodos de dissimilaridade, dos tipos de distância entre vetores e dos tipos de vizinhança foram analisadas. Por um lado, os resultados mostraram que os métodos de Ward e *complete linkage* tendem a fornecer *clusters* com tamanhos não tão desiguais. Por outro, os demais métodos tendem a formar *clusters* com tamanhos bastante diferentes. Isto sugere que, para fins de identificação de áreas geográficas homogêneas e relativamente próximas umas das outras, a utilização dos métodos de Ward e *complete linkage* pode ser mais aconselhável.

O fato de os métodos *single linkage*, *average linkage*, *average linkage weighted*, *centroid* e da mediana resultarem na formação de agrupamentos muito diferentes em termos de tamanho não necessariamente os exclui de aplicabilidade prática. A formação de uns poucos agrupamentos muito grandes *vis-à-vis* uma série de outros agrupamentos mais reduzidos pode ser utilizada como um instrumento para identificar conjuntos de poucos municípios que sejam significantemente diferentes dos demais em relação a determinadas variáveis de interesse. Portanto, estes cinco métodos podem ser utilizados como alternativas para os métodos *scan* (KULLDORFF, 1997; GLAZ e BALAKRISHNAN, 1999; GLAZ *et al.*, 2001), por exemplo.

Em relação à distribuição do número de municípios em cada agrupamento geográfico formado e ao comportamento dos critérios de seleção do número de

clusters, os resultados obtidos foram muito parecidos, utilizando-se tanto a vizinhança do tipo *queen* quanto a do tipo *rook*. Em termos de visualização geográfica, entretanto, a vizinhança do tipo *queen* permite uma maior flexibilidade na formação dos agrupamentos, uma vez que ela exige apenas um vértice em comum para caracterizar a vizinhança entre dois municípios. A vizinhança do tipo *rook*, por exigir um vértice em comum para caracterizar a vizinhança, possibilita a formação de agrupamentos com forma menos irregular. Dada a natureza hierárquica dos algoritmos de clusterização propostos neste estudo, não necessariamente os *clusters* formados com a vizinhança do tipo *queen* possuem menor variabilidade que os grupamentos formados com a vizinhança do tipo *rook*.

O estudo apresenta 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, são 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*, como uma medida de variabilidade total *intra-clusters*. Os resultados mostraram a capacidade do método *complete linkage*, e principalmente do método de Ward, de gerar agrupamentos com variabilidade menor que a dos agrupamentos políticos. No caso de microrregiões, por exemplo, o método de Ward possibilitou a formação de agrupamentos homogêneos, com metade da variabilidade dos agrupamentos políticos. Para os demais cinco métodos, devido à tendência de formação de *clusters* com tamanhos muito desiguais, a variabilidade total obtida resultou, em vários casos, maior que a variabilidade no caso dos agrupamentos políticos.

Várias questões estão em aberto para serem investigadas, podendo levar ao aprimoramento do método aqui proposto. Primeiramente, seria interessante estender as distâncias entre vetores para incorporar outros tipos de variáveis. As distâncias estudadas neste trabalho são mais apropriadas para variáveis contínuas. Podem-se modificar os procedimentos dos algoritmos de clusterização espacial hierárquica para tratar dados binários ou dados categóricos em geral. 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 propostos neste estudo são puramente heurísticos, e não estão baseados em modelos probabilísticos 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 apresentados neste estudo.

Finalmente, outra abertura para novos estudos é a seleção do número de *clusters*. Apresentou-se aqui um estudo para avaliar o comportamento de 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 se tratando de clusterização não espacial, muitas vezes o interesse está em se chegar a um número razoavelmente pequeno de agrupamentos (da ordem de dez); havendo 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 alguma política pública mais eficiente. Por este motivo, é interessante terem *clusters* não muito grandes, para evitar a existência de 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.

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. Vol. 14, n. 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, 1999.

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.

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.

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

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

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

LI, X. *et al.* **Storm clustering for data-driven weather forecasting**. Technical report, University of Alabama, 2008.

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.

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

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.

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.

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.

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.

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 infra-estrutura domiciliar nos municípios brasileiros. Livro: Diferenças Regionais no Brasil: Caracterização e Evolução nos Últimos Anos. Brasília: Ipea. (**no prelo**). 2006.

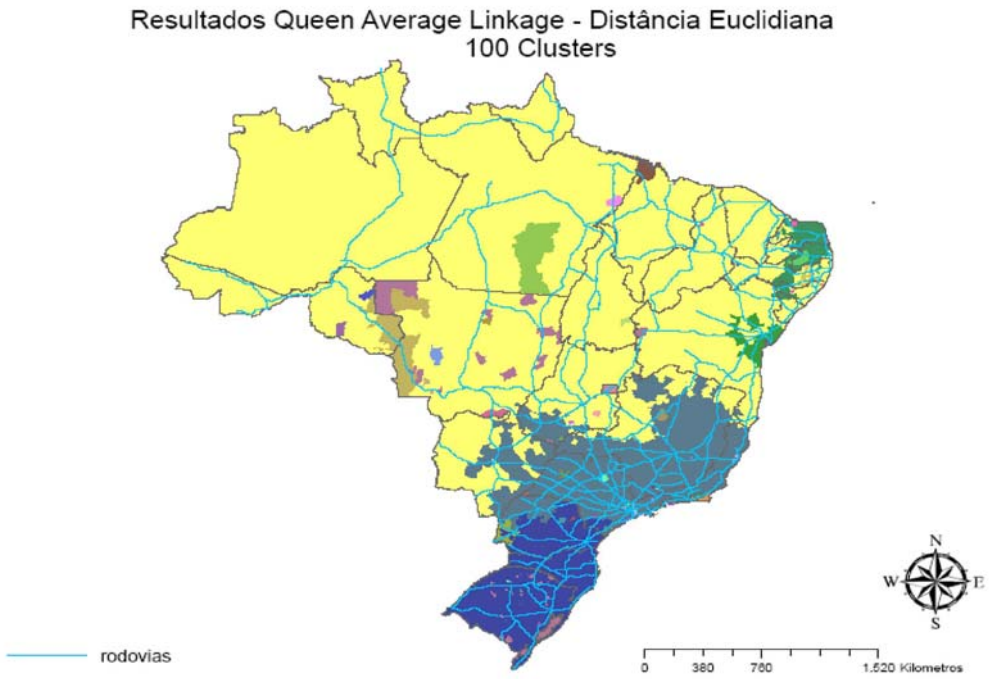
MILLIGAN, G. W.; COOPER, M. C. An examination of procedures for determining the number of clusters in a Data Set. **Psychometrika**, 50,159 – 179, 1985.

ANEXO 1

MAPAS DOS *CLUSTERS* FORMADOS COM OS DIFERENTES MÉTODOS E DIFERENTES DISTÂNCIAS[2]

Vizinhança do tipo *queen*

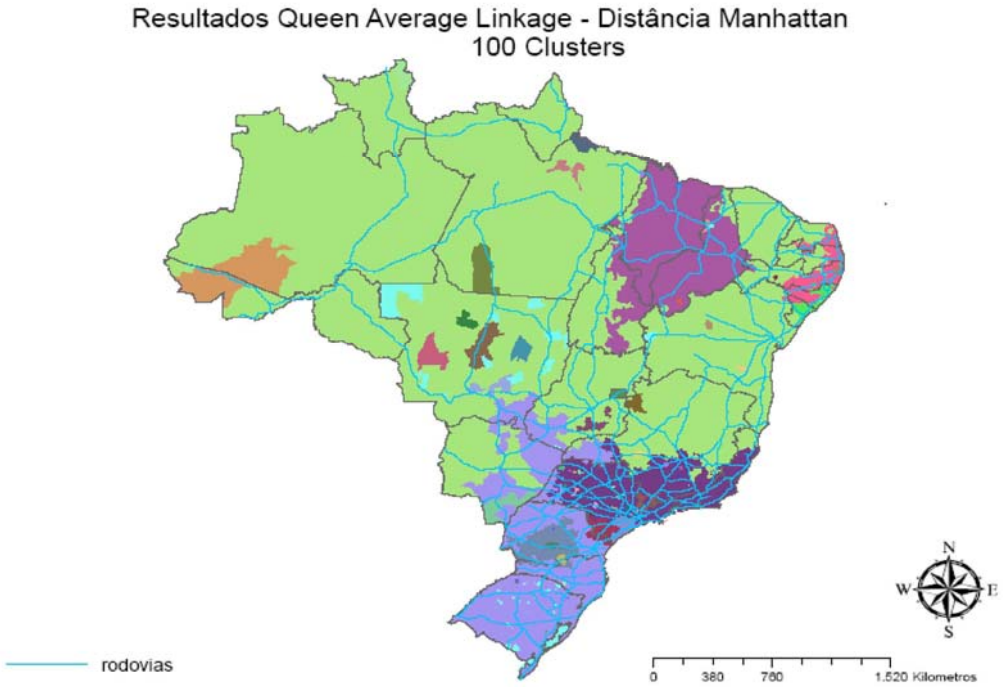
FIGURA A1.1
Mapa dos *clusters*



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).

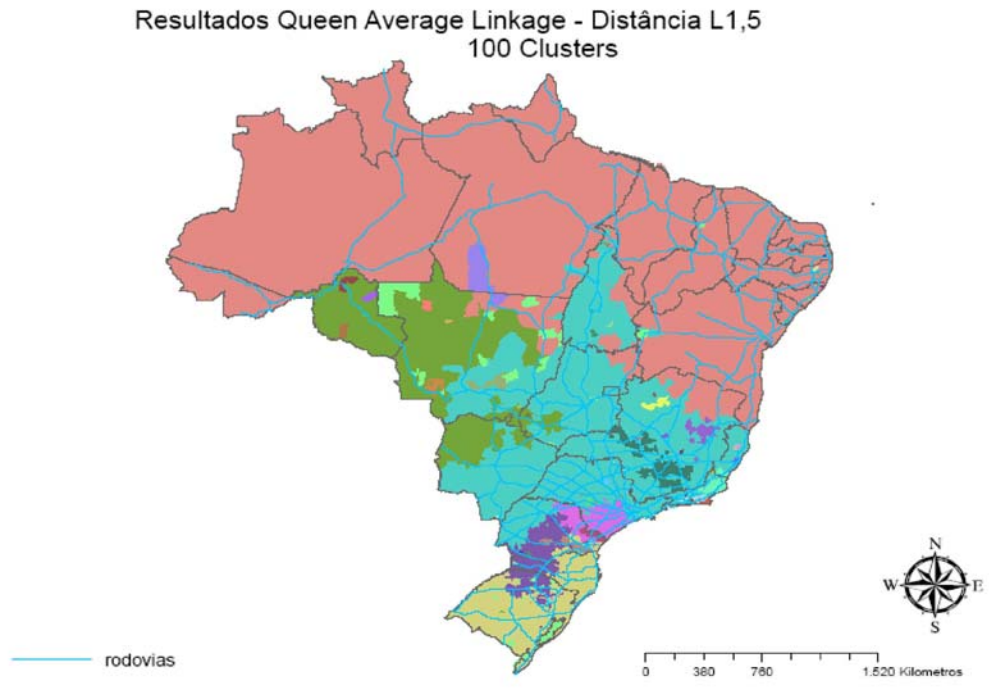
2. Para visualização em cores, acessar a seção *O trabalho do Ipea*, subseção *Publicações*, no site: <http://www.ipea.gov.br>.

FIGURA A1.2
Mapa dos *clusters*



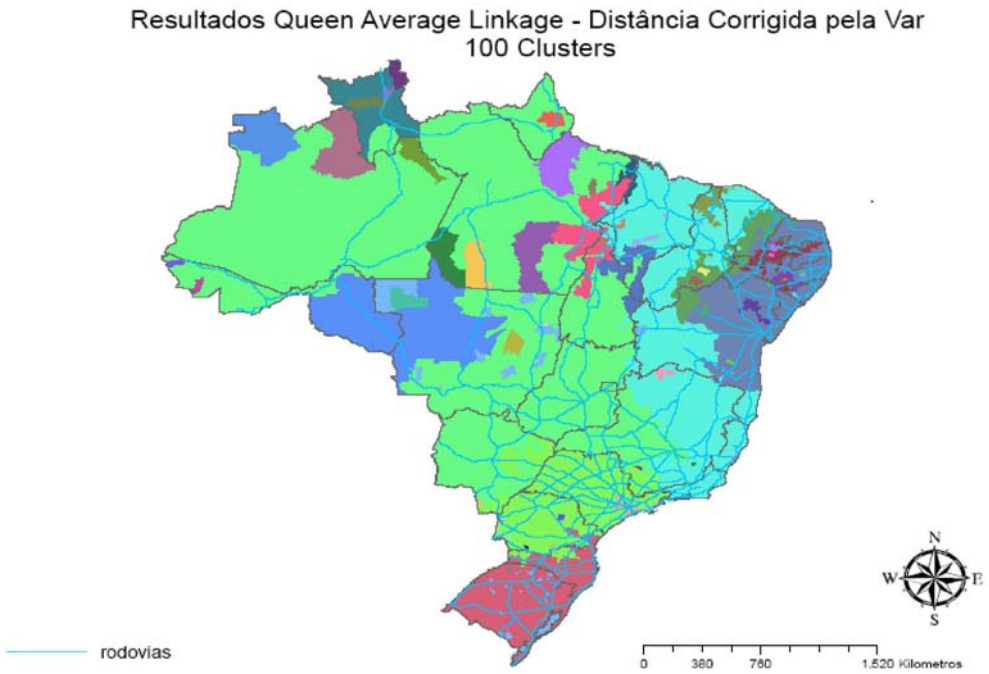
Elaboração dos autores.
Obs.: O mapa está reproduzido conforme o original fornecido pelos autores, cujas características não permitiram melhor ajuste para fim de impressão (nota do Editorial).

FIGURA A1.3
Mapa dos *clusters*



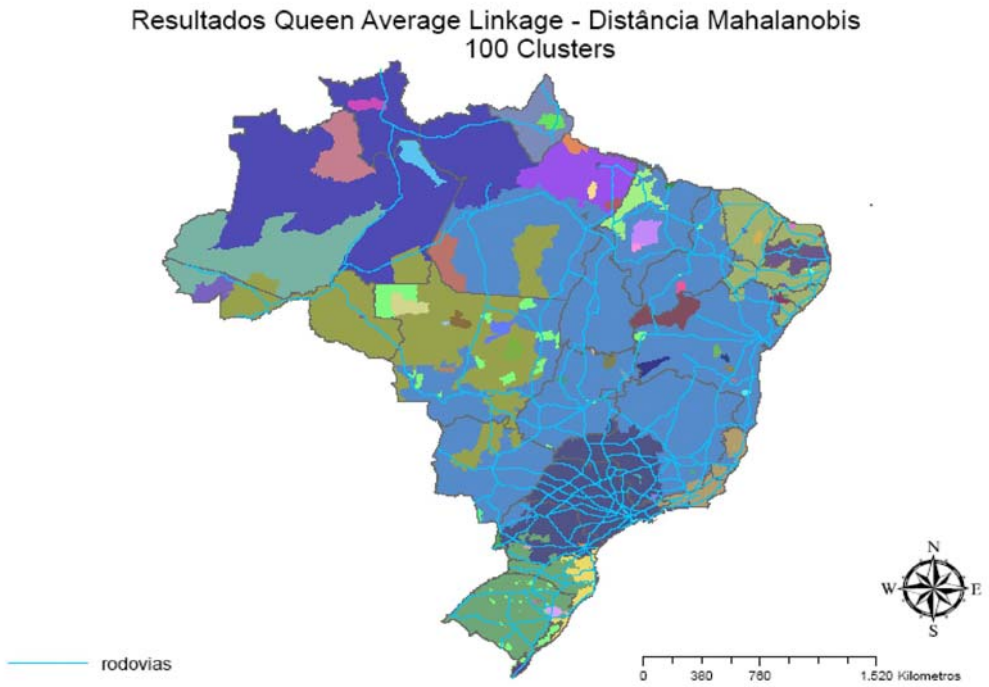
Elaboração dos autores.
Obs.: O mapa está reproduzido conforme o original fornecido pelos autores, cujas características não permitiram melhor ajuste para fim de impressão (nota do Editorial).

FIGURA A1.4
Mapa dos *clusters*



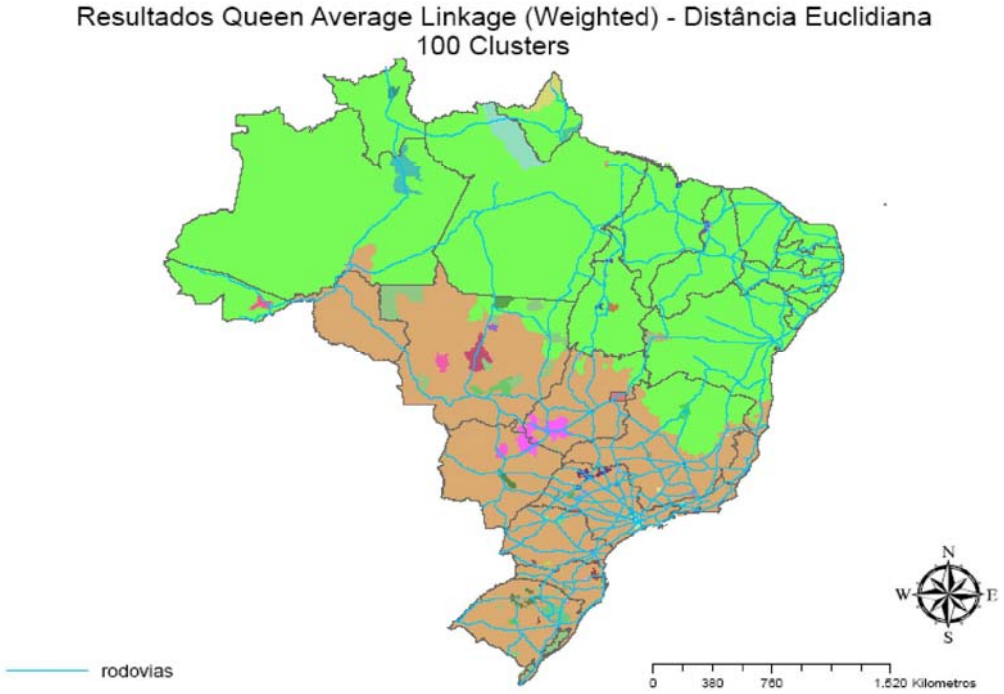
Elaboração dos autores.
Obs.: O mapa está reproduzido conforme o original fornecido pelos autores, cujas características não permitiram melhor ajuste para fim de impressão (nota do Editorial).

FIGURA A1.5
Mapa dos *clusters*



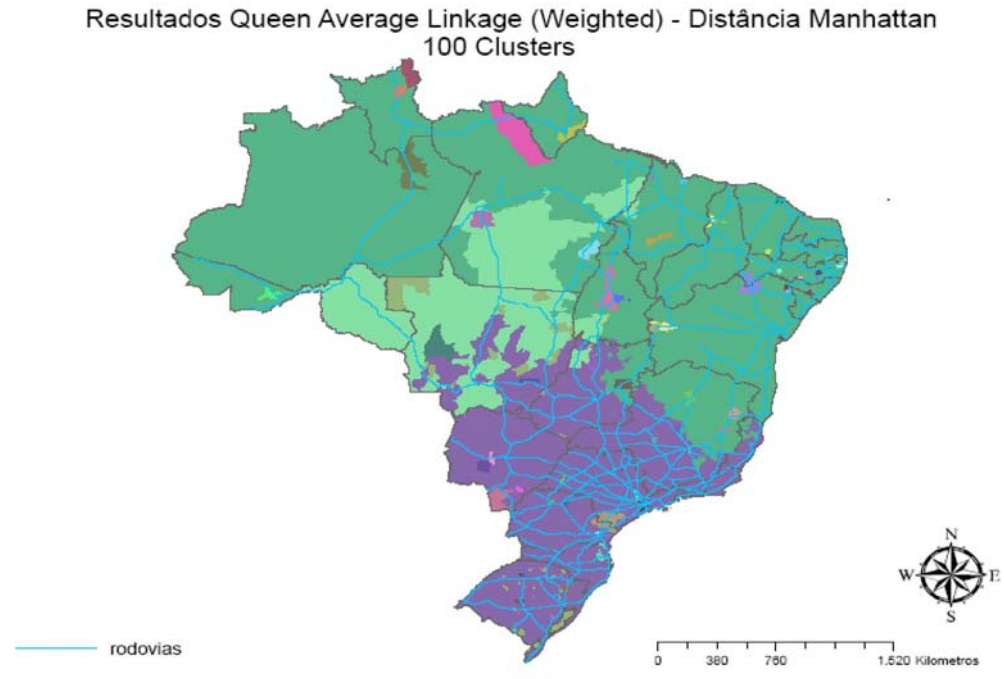
Elaboração dos autores.
Obs.: O mapa está reproduzido conforme o original fornecido pelos autores, cujas características não permitiram melhor ajuste para fim de impressão (nota do Editorial).

FIGURA A1.6
Mapa dos *clusters*



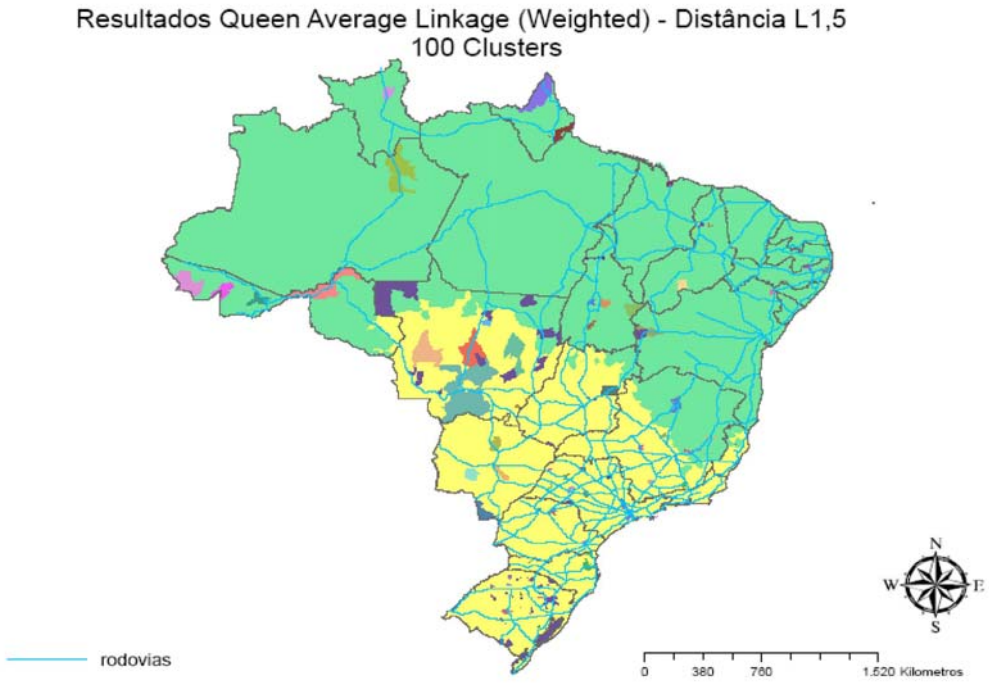
Elaboração dos autores.
Obs.: O mapa está reproduzido conforme o original fornecido pelos autores, cujas características não permitiram melhor ajuste para fim de impressão (nota do Editorial).

FIGURA A1.7
Mapa dos *clusters*



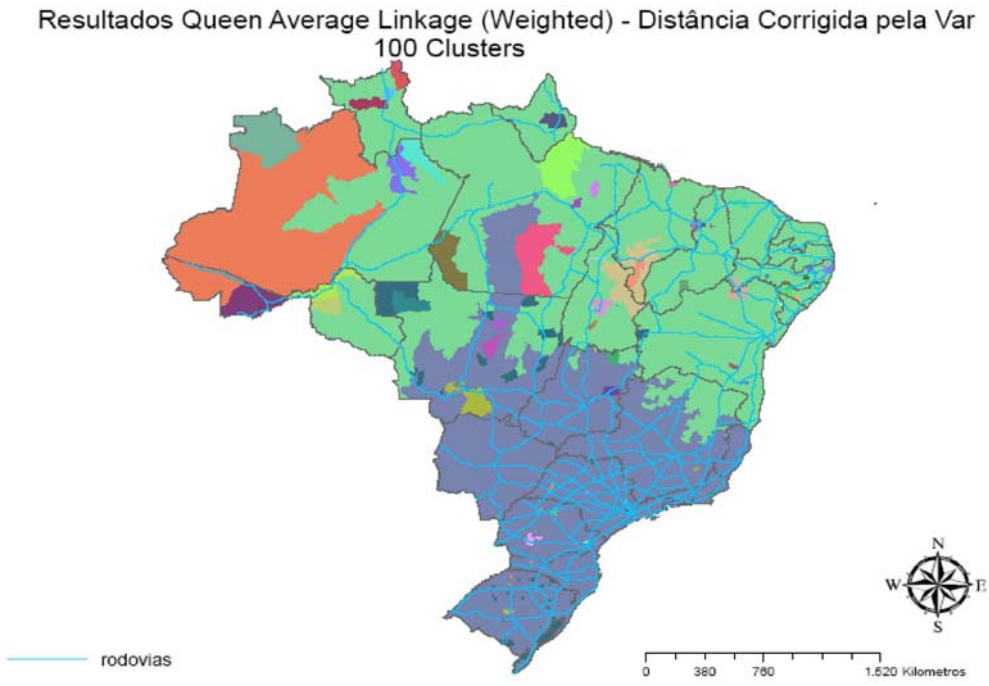
Elaboração dos autores.
Obs.: O mapa está reproduzido conforme o original fornecido pelos autores, cujas características não permitiram melhor ajuste para fim de impressão (nota do Editorial).

FIGURA A1.8
Mapa dos *clusters*



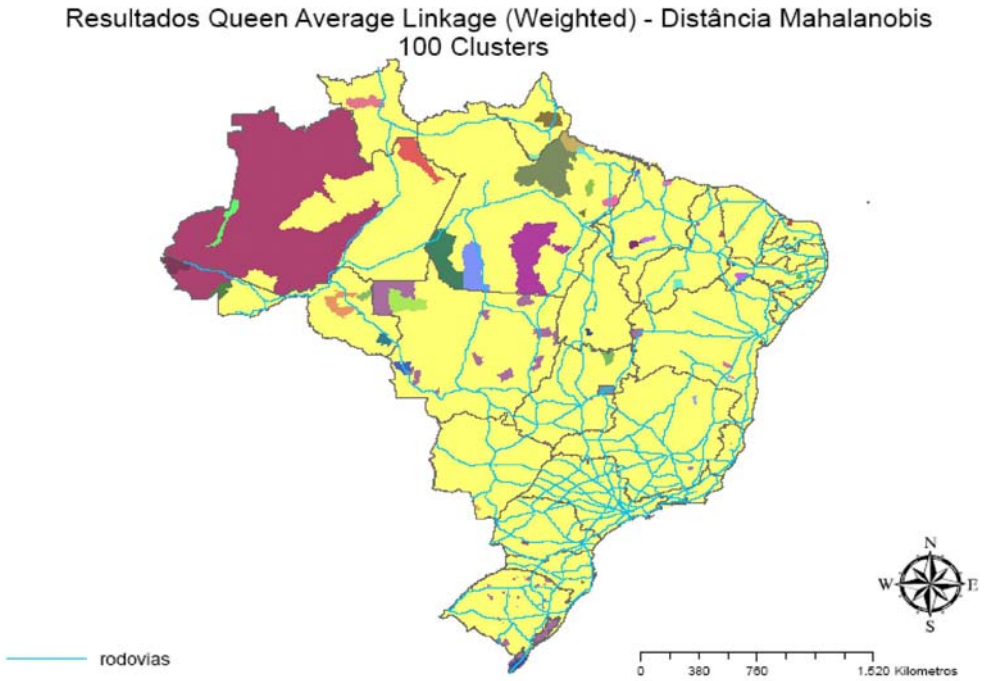
Elaboração dos autores.
Obs.: O mapa está reproduzido conforme o original fornecido pelos autores, cujas características não permitiram melhor ajuste para fim de impressão (nota do Editorial).

FIGURA A1.9
Mapa dos *clusters*



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).

FIGURA A1.10
Mapa dos *clusters*

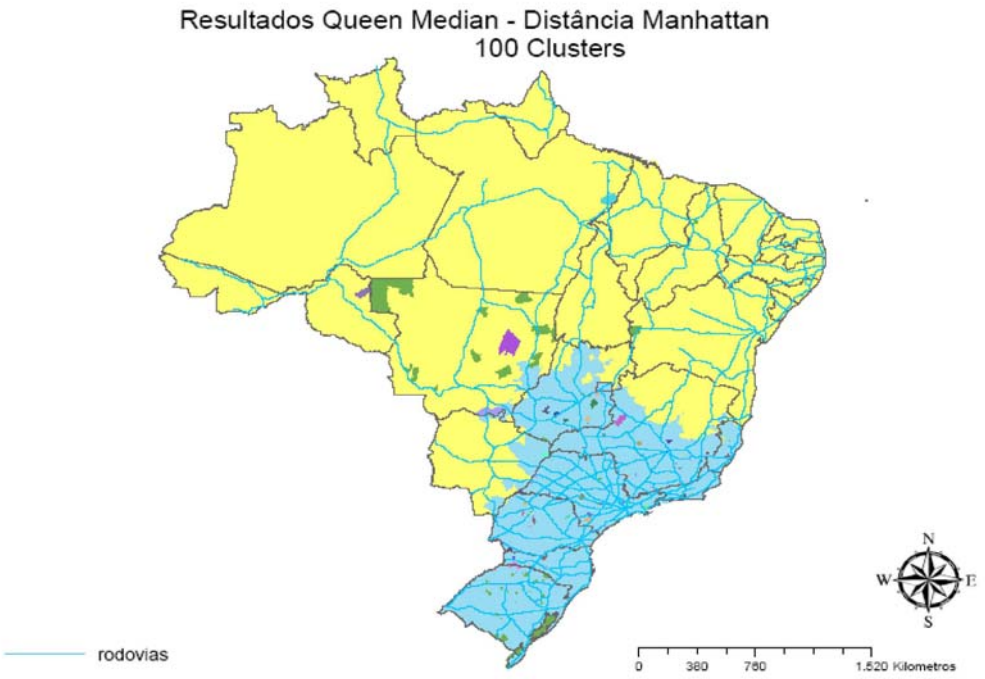


Elaboração dos autores.
Obs.: O mapa está reproduzido conforme o original fornecido pelos autores, cujas características não permitiram melhor ajuste para fim de impressão (nota do Editorial).

FIGURA A1.11
Mapa dos *clusters*

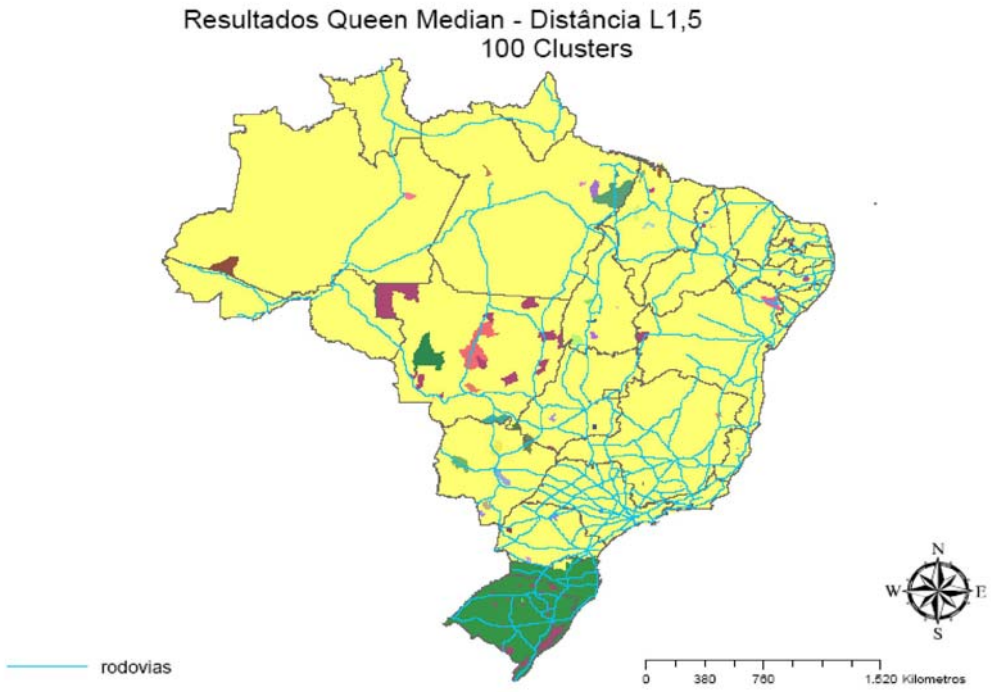
Elaboração dos autores.
Obs.: O mapa está reproduzido conforme o original fornecido pelos autores, cujas características não permitiram melhor ajuste para fim de impressão (nota do Editorial).

FIGURA A1.12
Mapa dos *clusters*



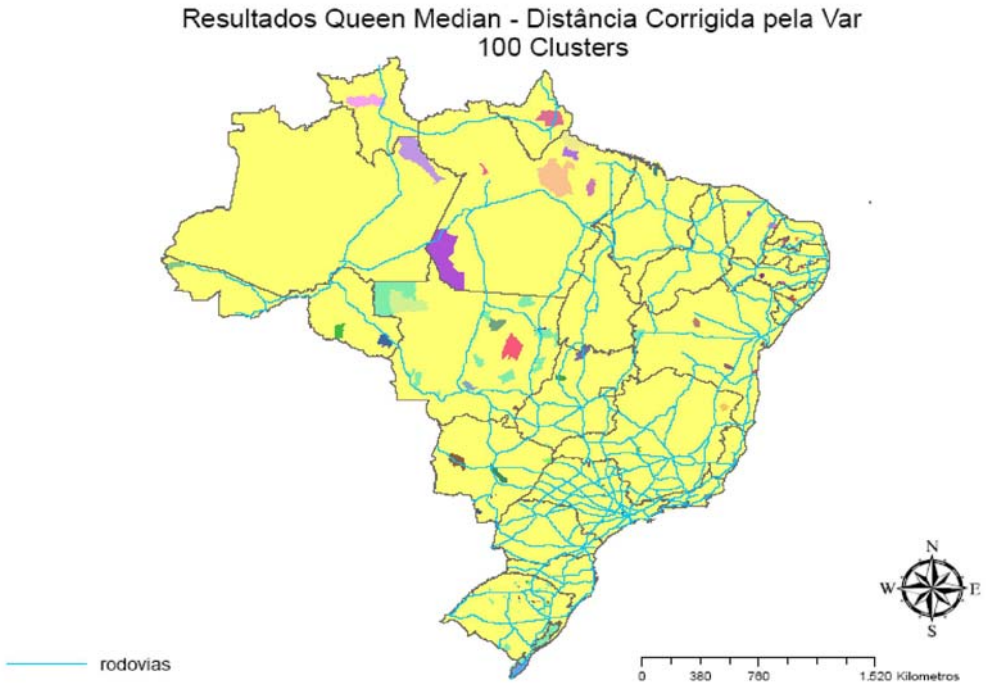
Elaboração dos autores.
Obs.: O mapa está reproduzido conforme o original fornecido pelos autores, cujas características não permitiram melhor ajuste para fim de impressão (nota do Editorial).

FIGURA A1.13
Mapa dos *clusters*



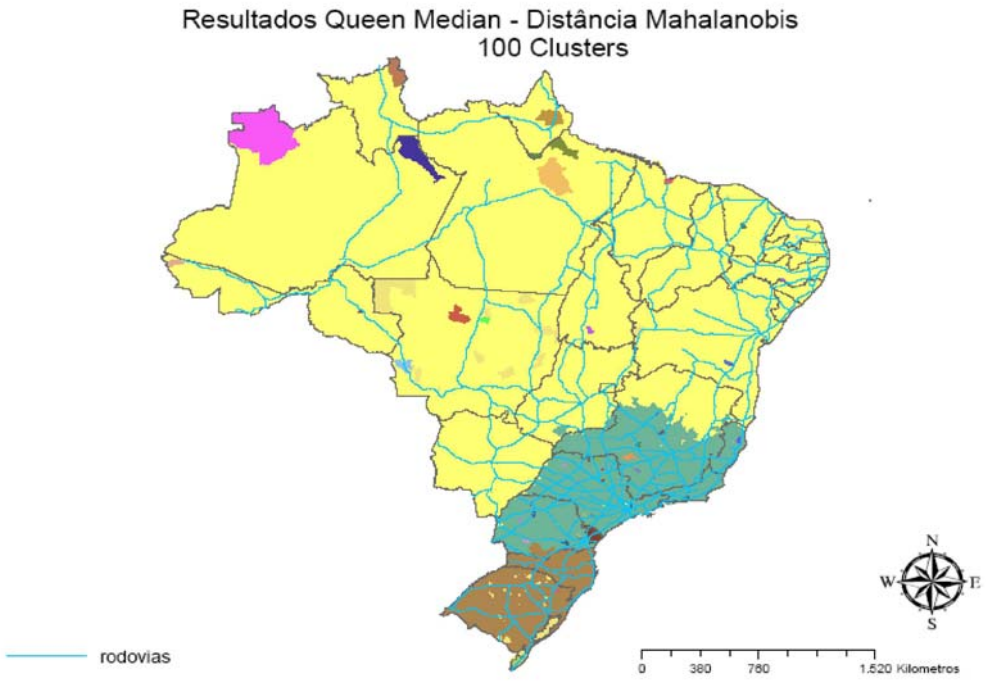
Elaboração dos autores.
Obs.: O mapa está reproduzido conforme o original fornecido pelos autores, cujas características não permitiram melhor ajuste para fim de impressão (nota do Editorial).

FIGURA A1.14
Mapa dos *clusters*



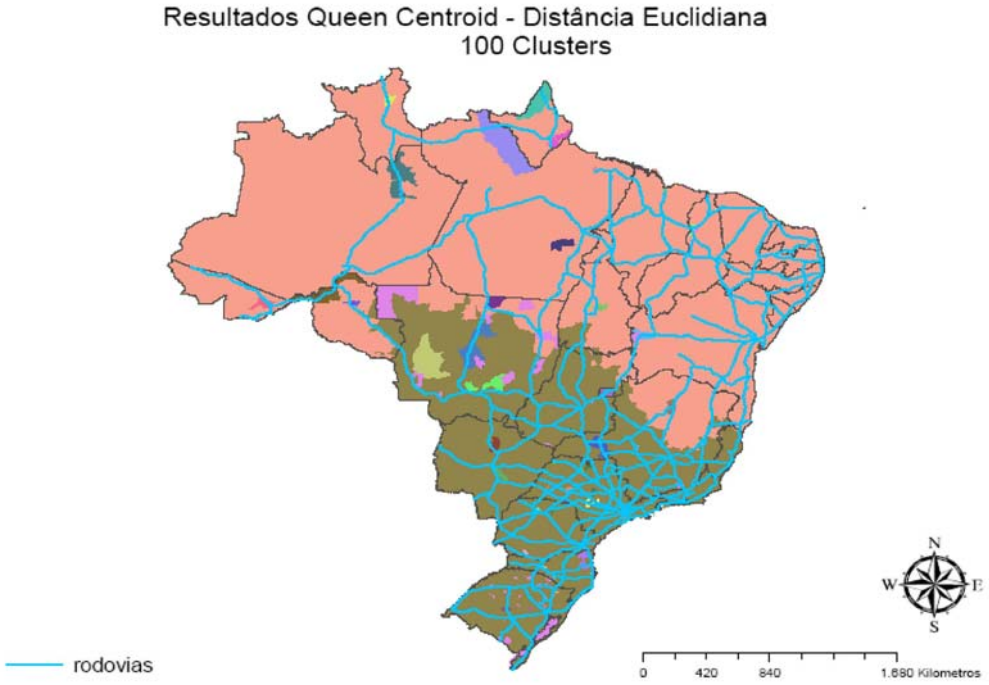
Elaboração dos autores.
Obs.: O mapa está reproduzido conforme o original fornecido pelos autores, cujas características não permitiram melhor ajuste para fim de impressão (nota do Editorial).

FIGURA A1.15
Mapa dos *clusters*



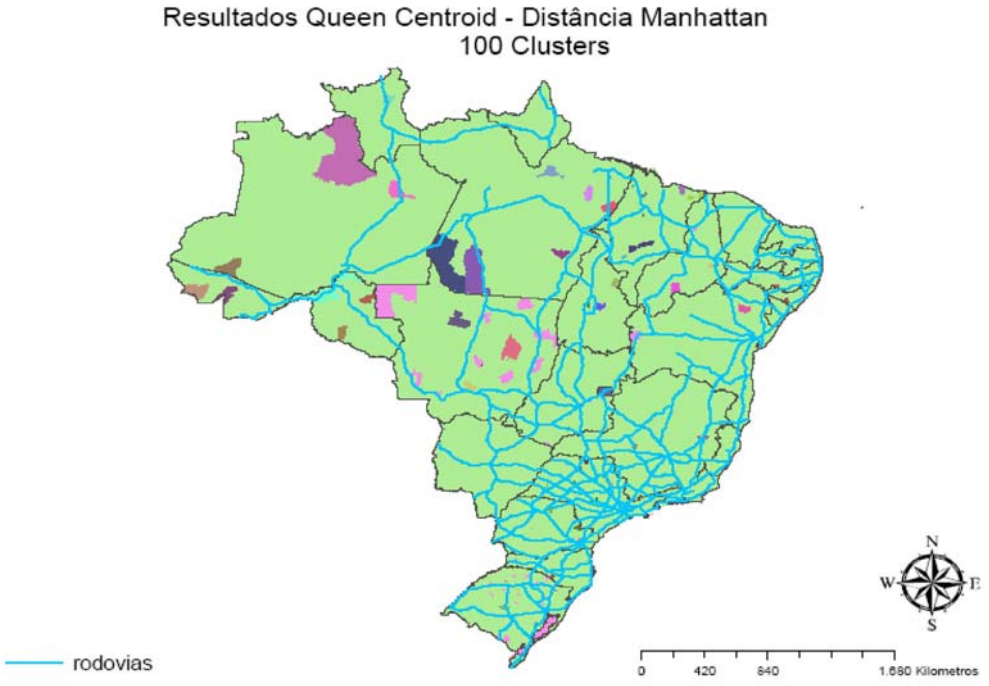
Elaboração dos autores.
Obs.: O mapa está reproduzido conforme o original fornecido pelos autores, cujas características não permitiram melhor ajuste para fim de impressão (nota do Editorial).

FIGURA A1.16
Mapa dos *clusters*



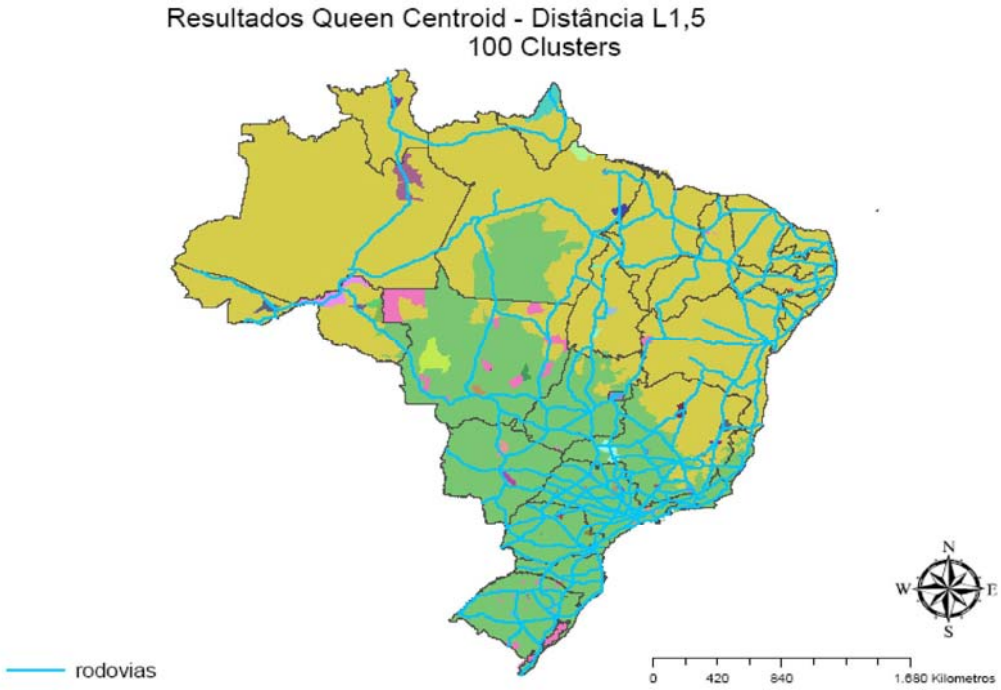
Elaboração dos autores.
Obs.: O mapa está reproduzido conforme o original fornecido pelos autores, cujas características não permitiram melhor ajuste para fim de impressão (nota do Editorial).

FIGURA A1.17
Mapa dos *clusters*



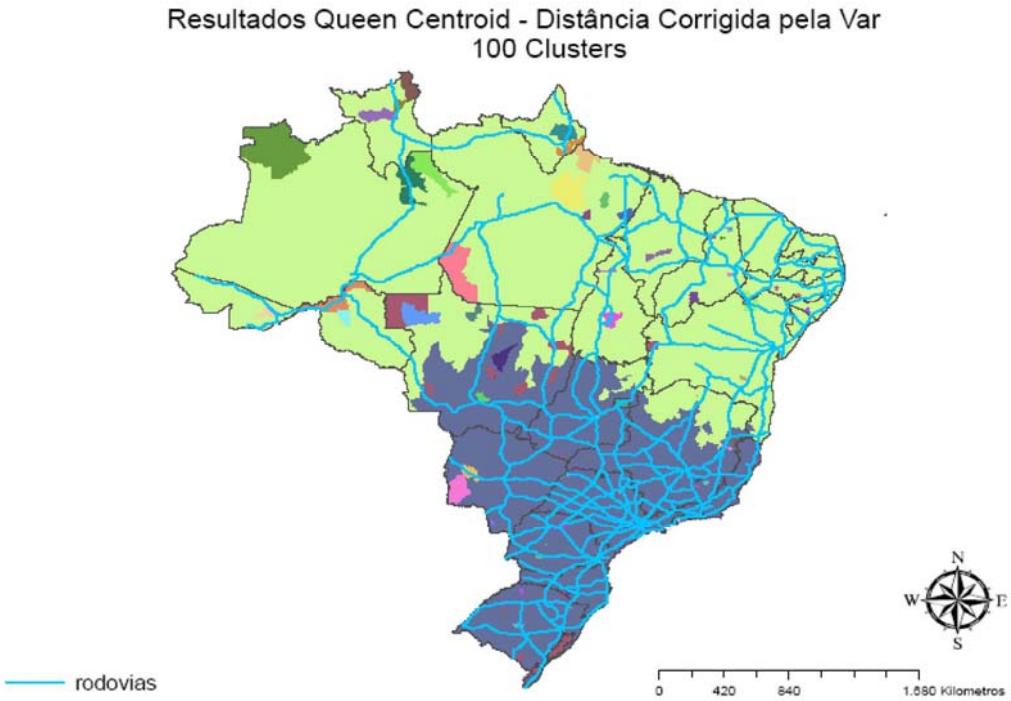
Elaboração dos autores.
Obs.: O mapa está reproduzido conforme o original fornecido pelos autores, cujas características não permitiram melhor ajuste para fim de impressão (nota do Editorial).

FIGURA A1.18
Mapa dos *clusters*



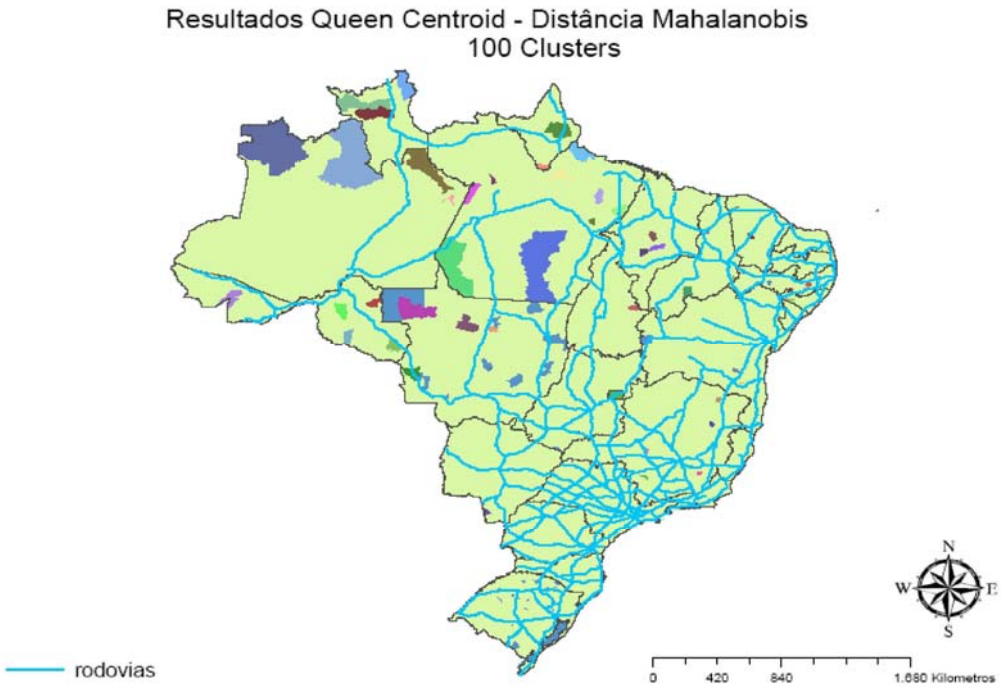
Elaboração dos autores.
Obs.: O mapa está reproduzido conforme o original fornecido pelos autores, cujas características não permitiram melhor ajuste para fim de impressão (nota do Editorial).

FIGURA A1.19
Mapa dos *clusters*



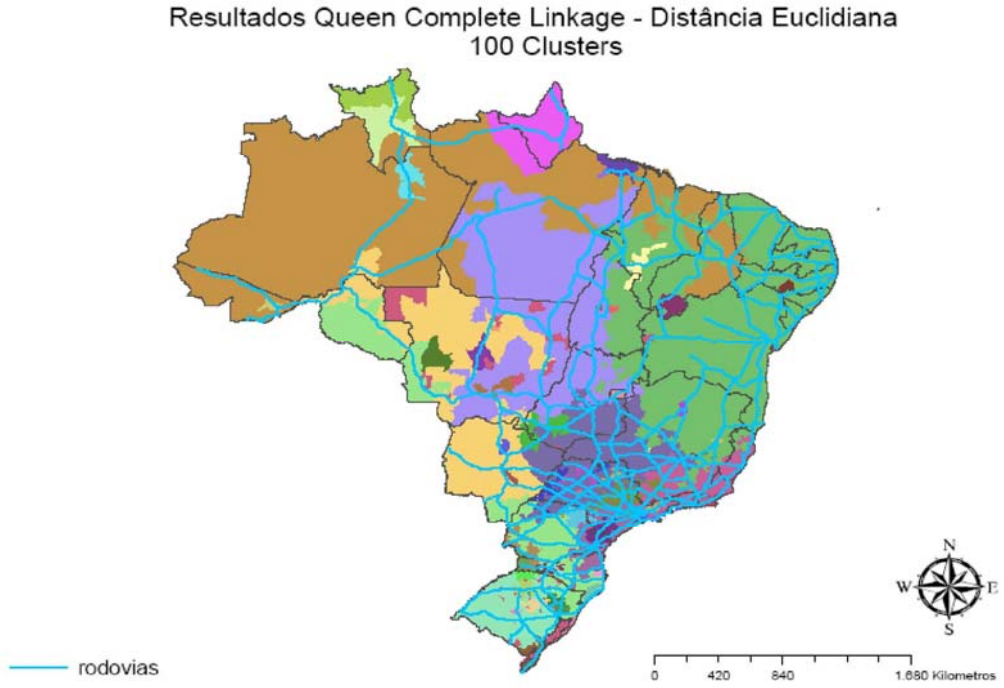
Elaboração dos autores.
Obs.: O mapa está reproduzido conforme o original fornecido pelos autores, cujas características não permitiram melhor ajuste para fim de impressão (nota do Editorial).

FIGURA A1.20
Mapa dos *clusters*



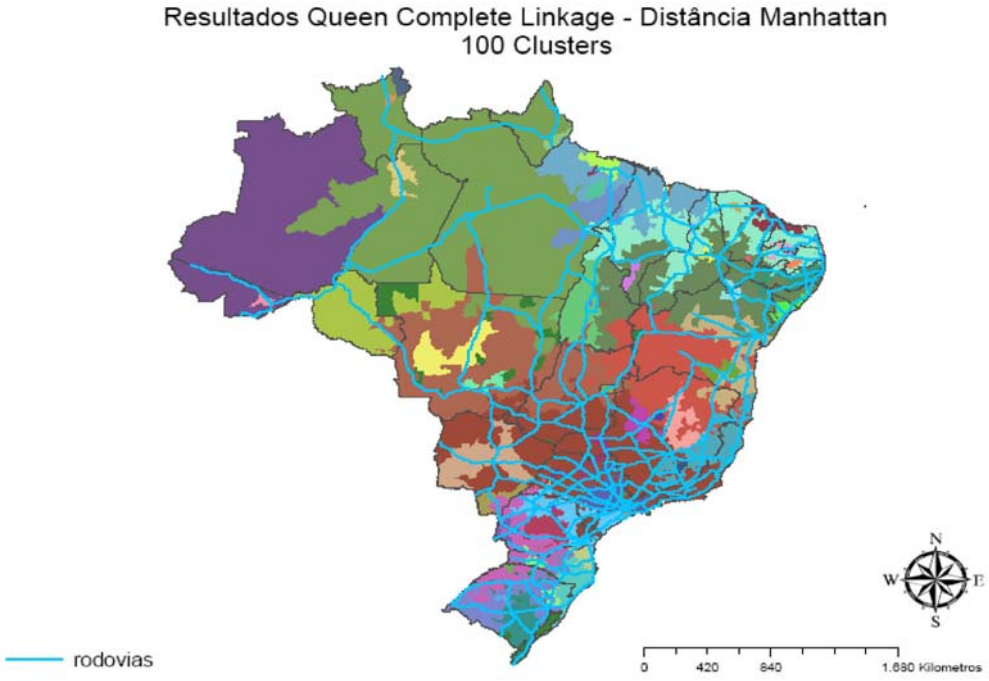
Elaboração dos autores.
Obs.: O mapa está reproduzido conforme o original fornecido pelos autores, cujas características não permitiram melhor ajuste para fim de impressão (nota do Editorial).

FIGURA A1.21
Mapa dos *clusters*



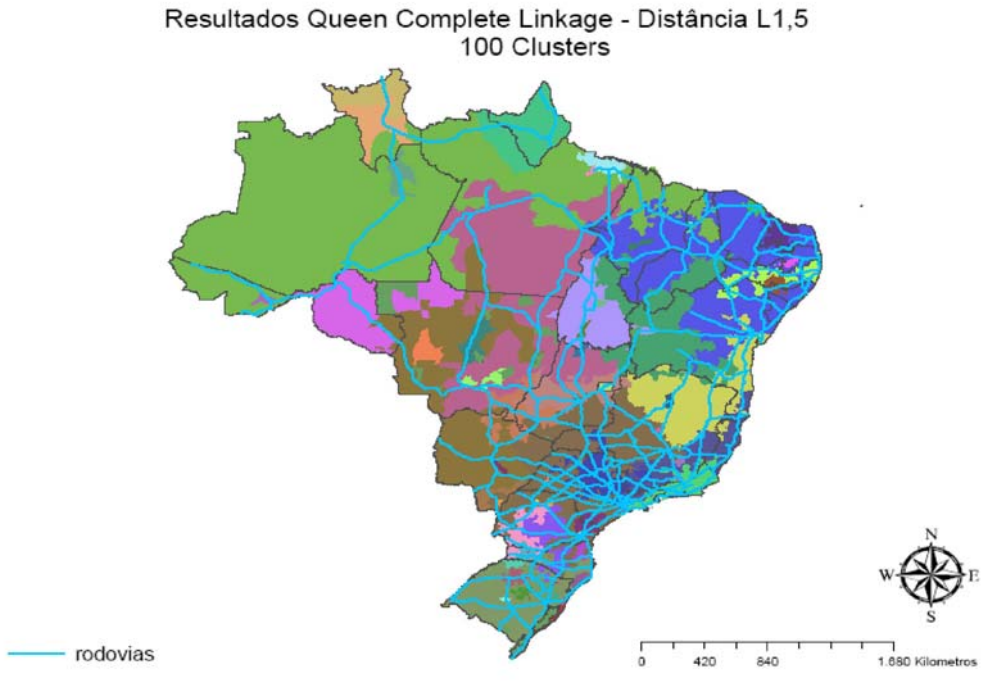
Elaboração dos autores.
Obs.: O mapa está reproduzido conforme o original fornecido pelos autores, cujas características não permitiram melhor ajuste para fim de impressão (nota do Editorial).

FIGURA A1.22
Mapa dos *clusters*



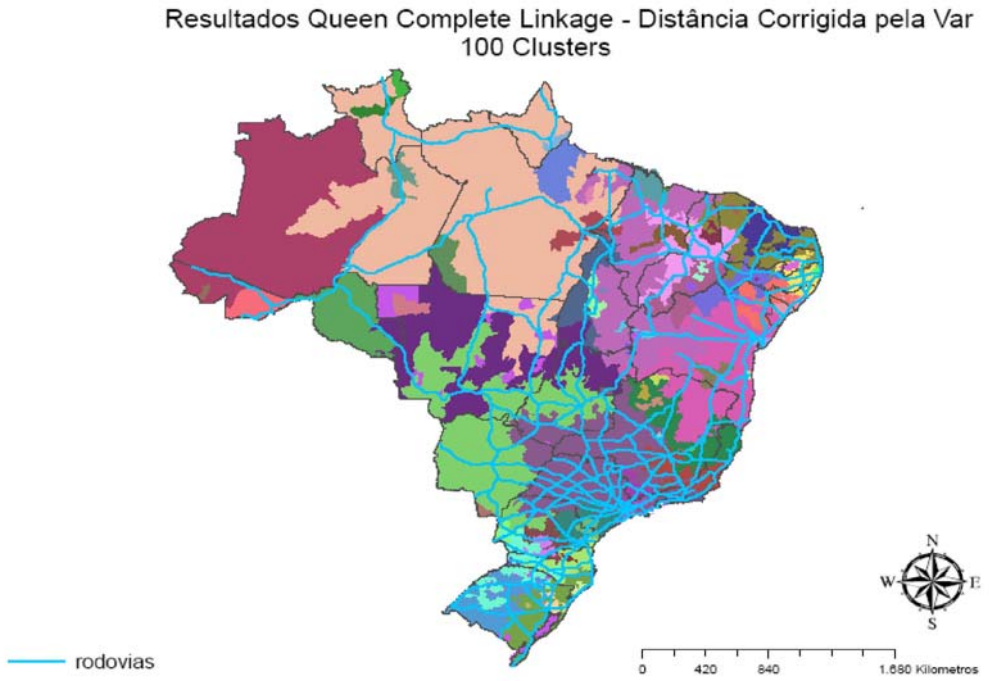
Elaboração dos autores.
Obs.: O mapa está reproduzido conforme o original fornecido pelos autores, cujas características não permitiram melhor ajuste para fim de impressão (nota do Editorial).

FIGURA A1.23
Mapa dos *clusters*



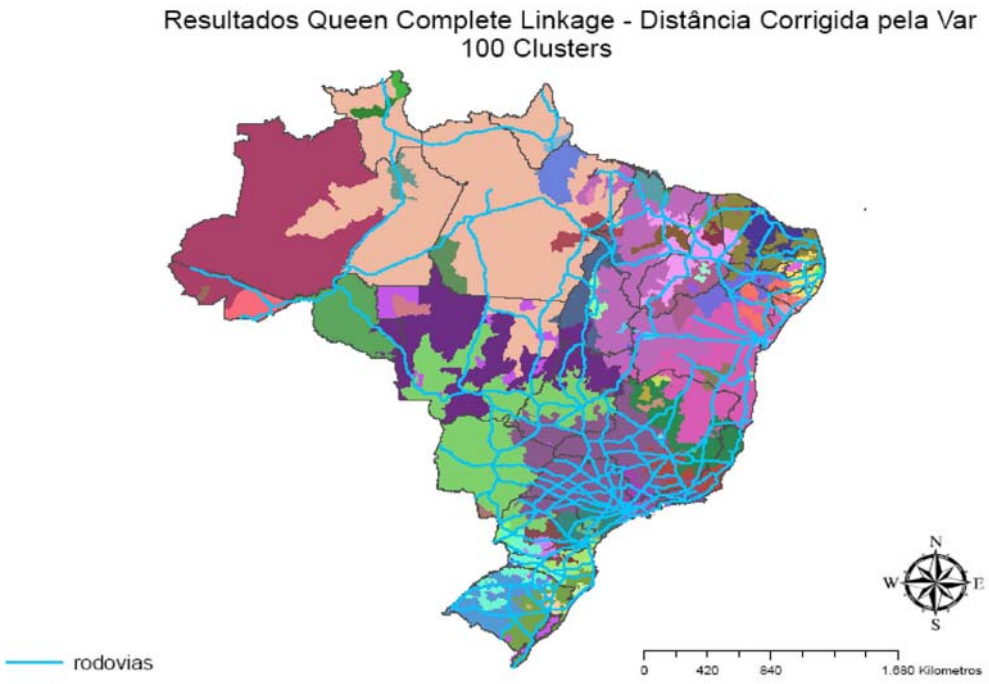
Elaboração dos autores.
Obs.: O mapa está reproduzido conforme o original fornecido pelos autores, cujas características não permitiram melhor ajuste para fim de impressão (nota do Editorial).

FIGURA A1.24
Mapa dos *clusters*



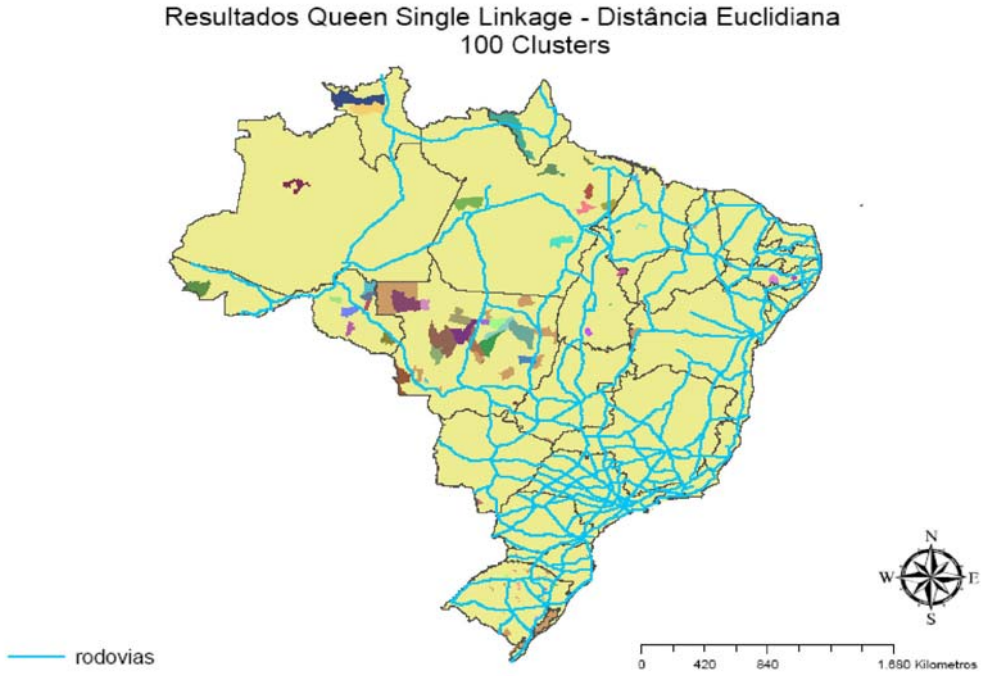
Elaboração dos autores.
Obs.: O mapa está reproduzido conforme o original fornecido pelos autores, cujas características não permitiram melhor ajuste para fim de impressão (nota do Editorial).

FIGURA A1.25
Mapa dos *clusters*



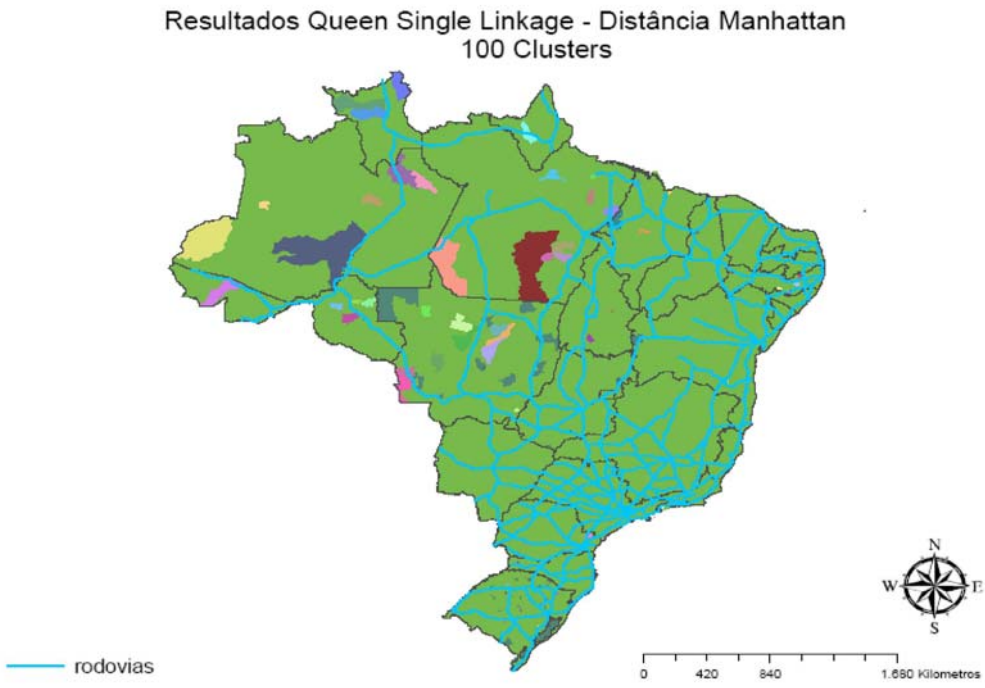
Elaboração dos autores.
Obs.: O mapa está reproduzido conforme o original fornecido pelos autores, cujas características não permitiram melhor ajuste para fim de impressão (nota do Editorial).

FIGURA A1.26
Mapa dos *clusters*



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).

FIGURA A1.27
Mapa dos *clusters*

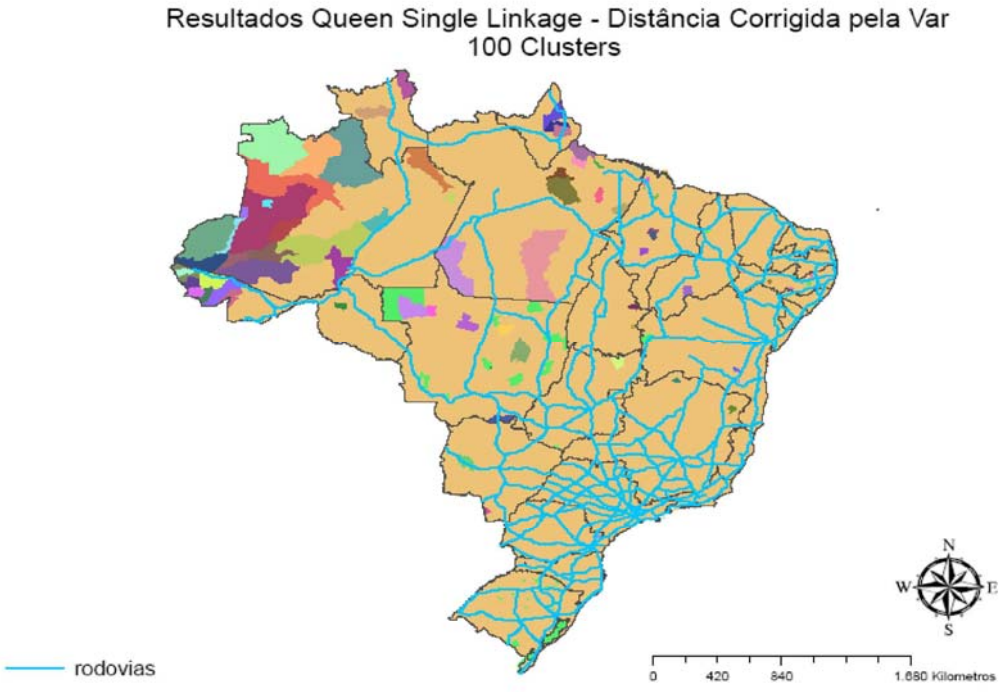


Elaboração dos autores.
Obs.: O mapa está reproduzido conforme o original fornecido pelos autores, cujas características não permitiram melhor ajuste para fim de impressão (nota do Editorial).

FIGURA A1.28
Mapa dos *clusters*

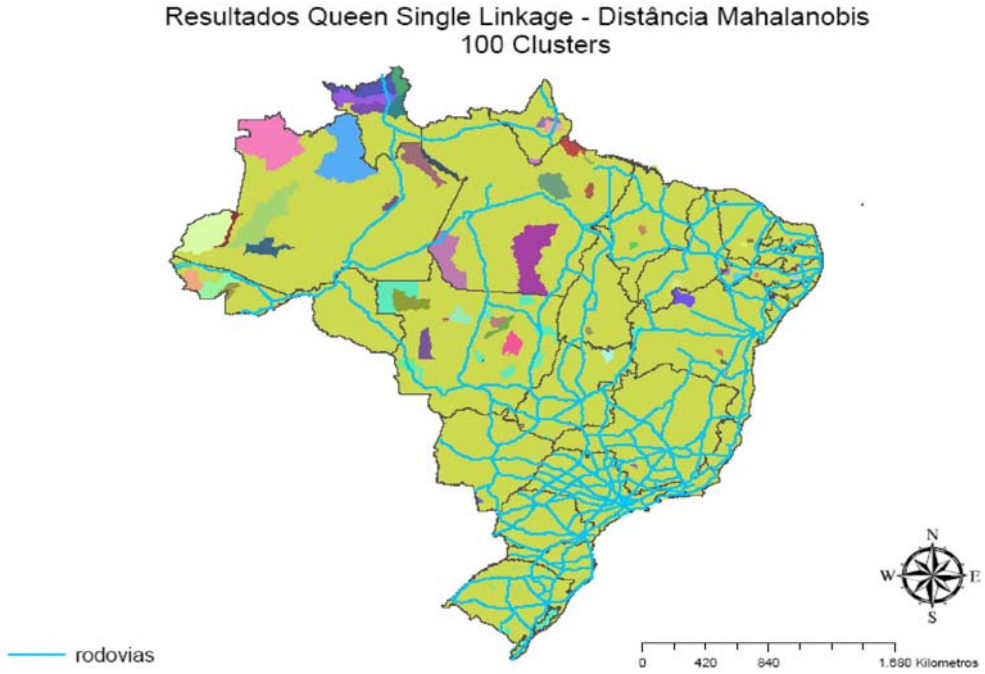
Elaboração dos autores.
Obs.: O mapa está reproduzido conforme o original fornecido pelos autores, cujas características não permitiram melhor ajuste para fim de impressão (nota do Editorial).

FIGURA A1.29
Mapa dos *clusters*



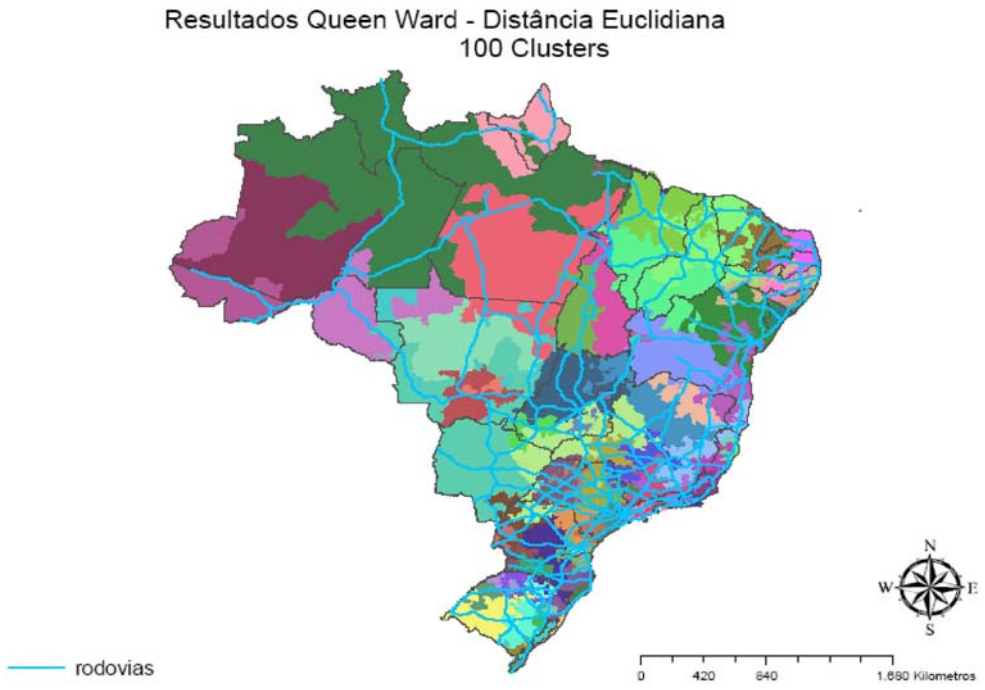
Elaboração dos autores.
Obs.: O mapa está reproduzido conforme o original fornecido pelos autores, cujas características não permitiram melhor ajuste para fim de impressão (nota do Editorial).

FIGURA A1.30
Mapa dos *clusters*



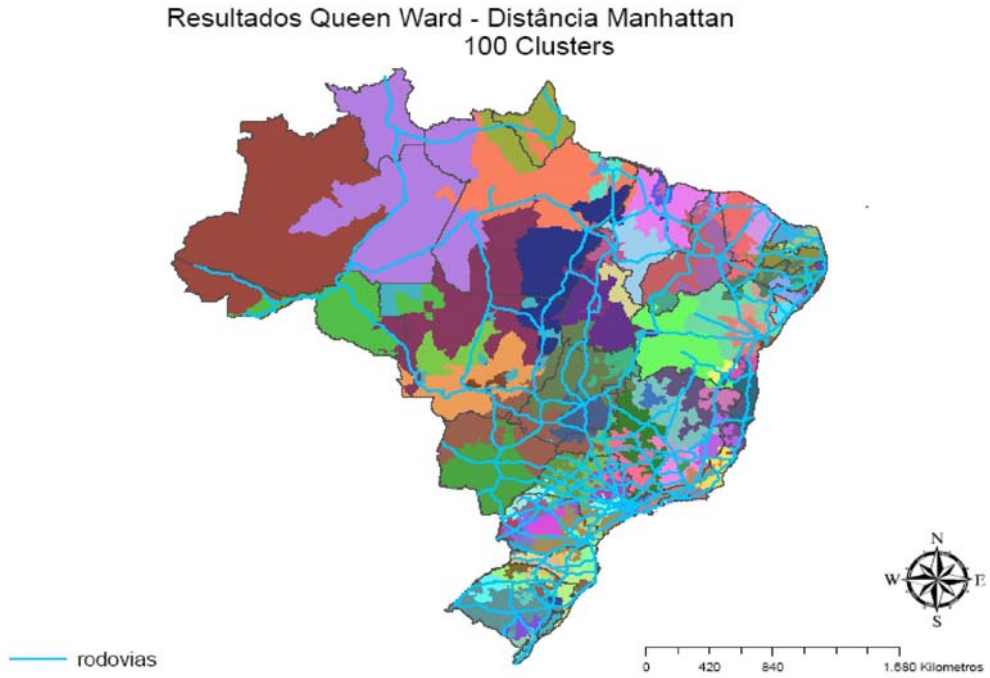
Elaboração dos autores.
Obs.: O mapa está reproduzido conforme o original fornecido pelos autores, cujas características não permitiram melhor ajuste para fim de impressão (nota do Editorial).

FIGURA A1.31
Mapa dos *clusters*



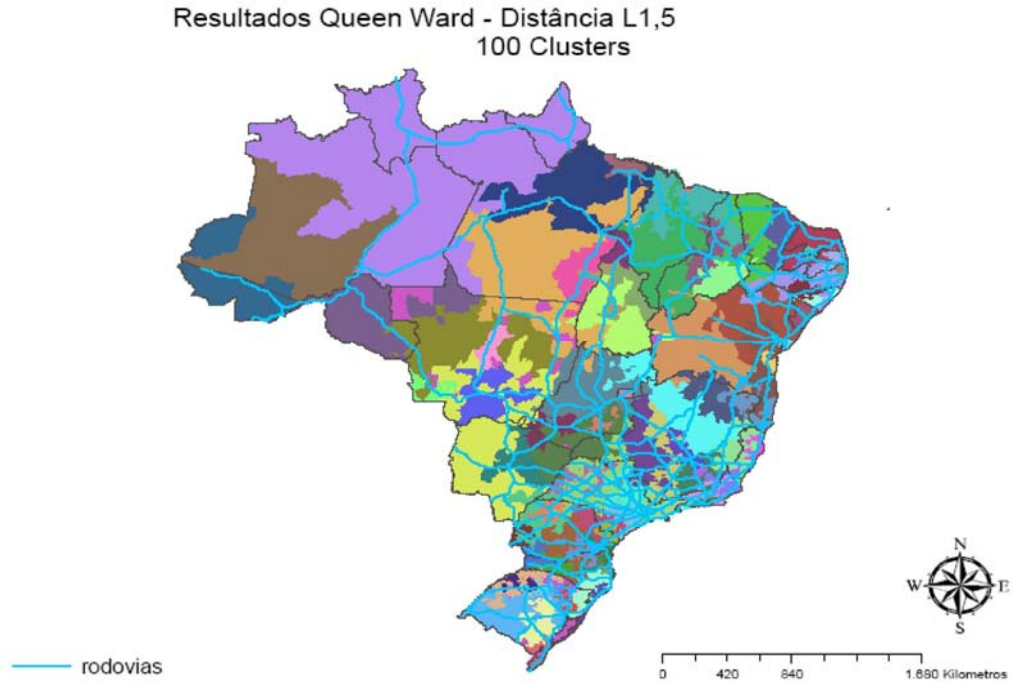
Elaboração dos autores.
Obs.: O mapa está reproduzido conforme o original fornecido pelos autores, cujas características não permitiram melhor ajuste para fim de impressão (nota do Editorial).

FIGURA A1.32
Mapa dos *clusters*



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).

FIGURA A1.33
Mapa dos *clusters*

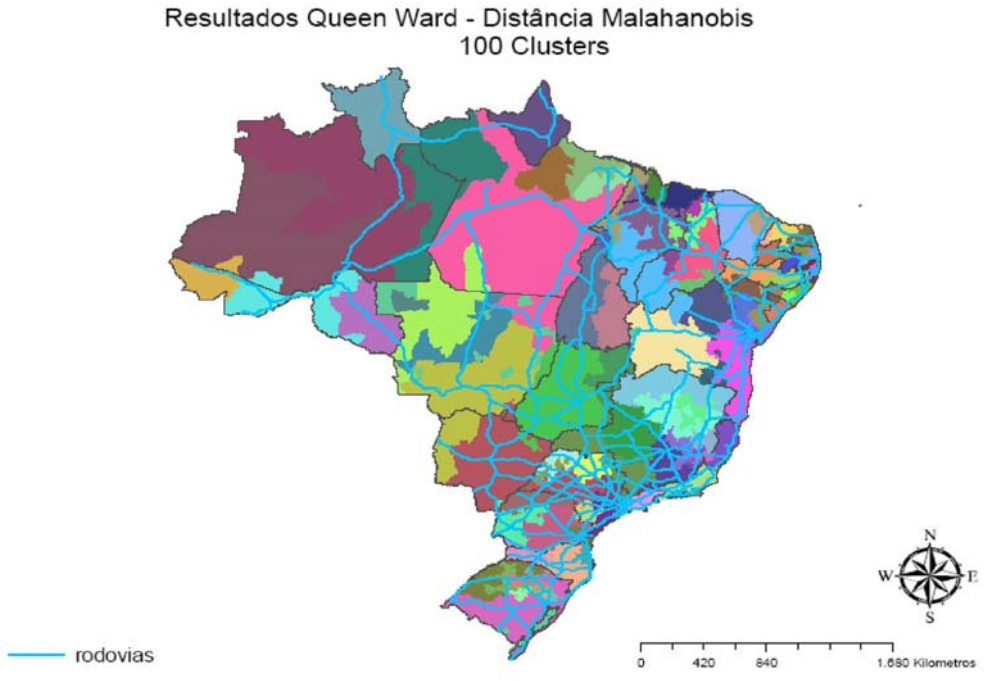


Elaboração dos autores.
Obs.: O mapa está reproduzido conforme o original fornecido pelos autores, cujas características não permitiram melhor ajuste para fim de impressão (nota do Editorial).

FIGURA A1.34
Mapa dos *clusters*

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).

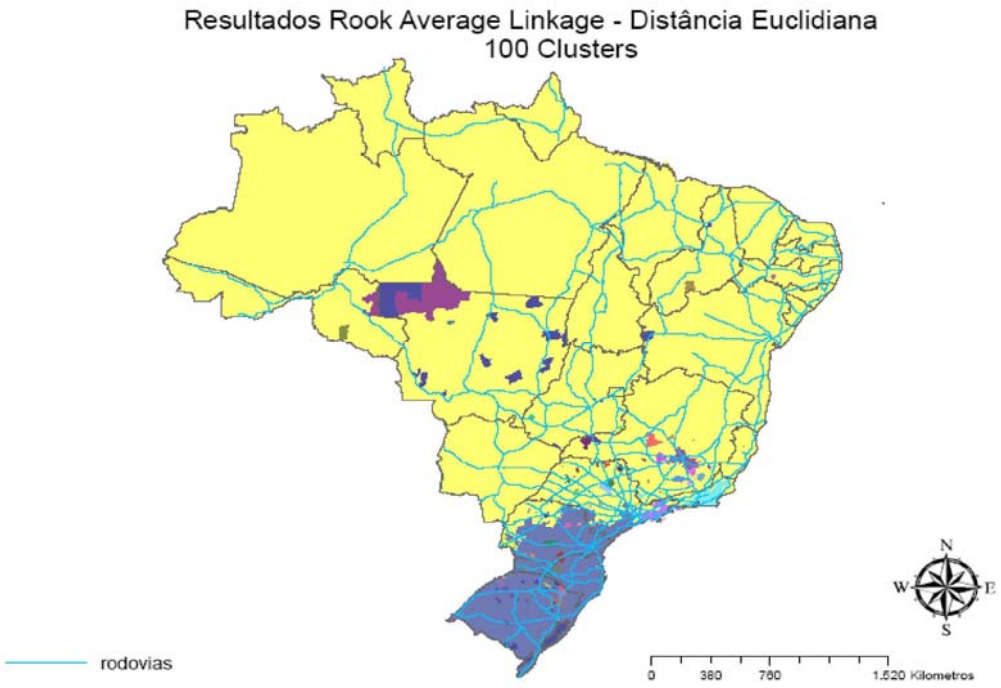
FIGURA A1.35
Mapa dos *clusters*



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).

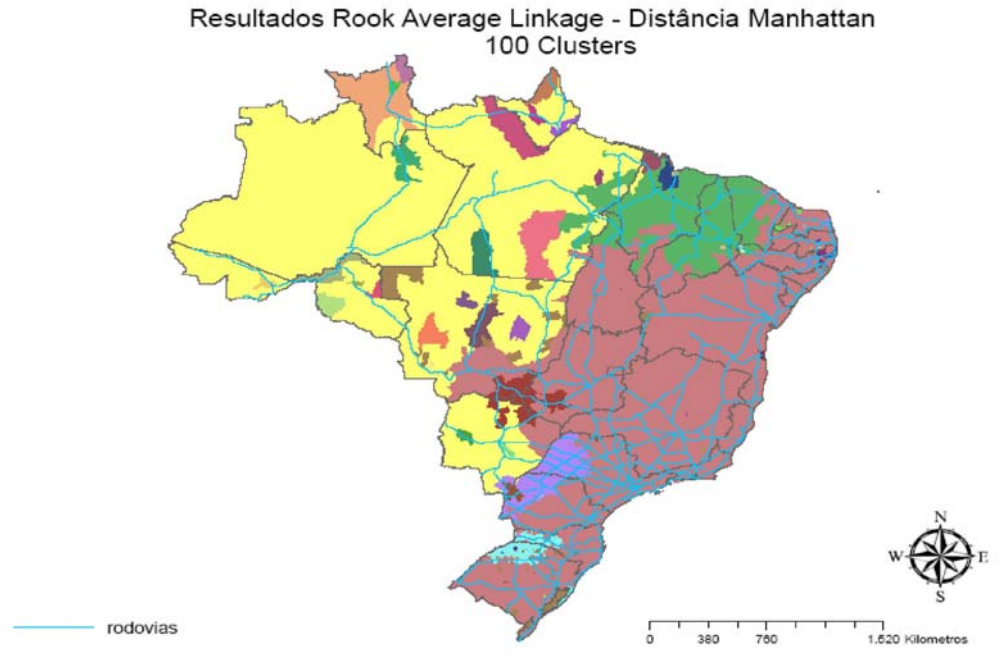
Vizinhança do tipo *rook*

FIGURA A1.36
Mapa dos *clusters*



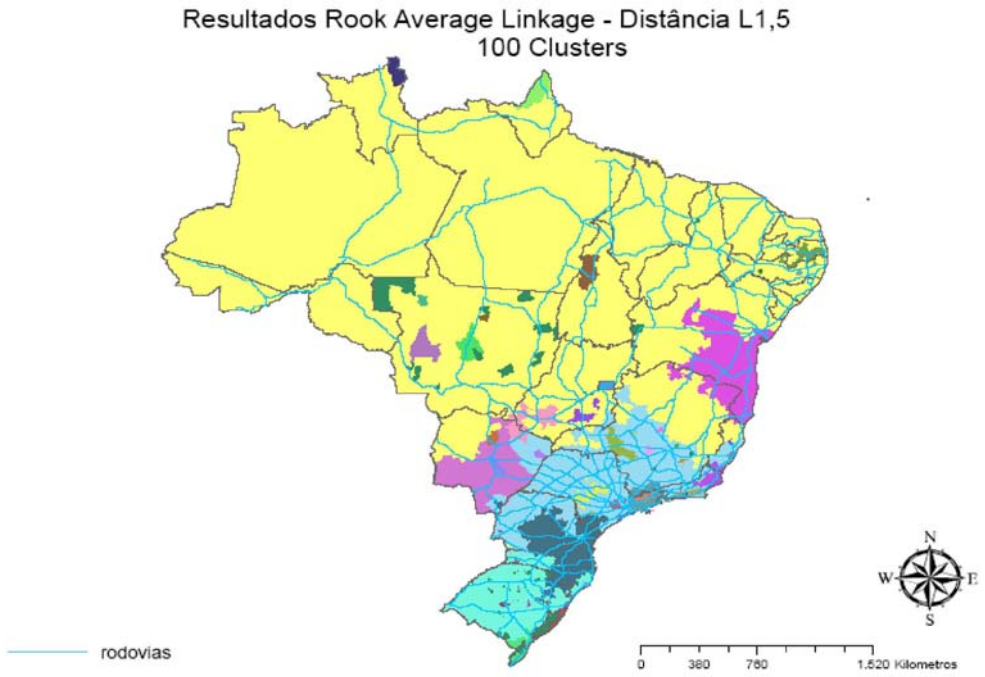
Elaboração dos autores.
Obs.: O mapa está reproduzido conforme o original fornecido pelos autores, cujas características não permitiram melhor ajuste para fim de impressão (nota do Editorial).

FIGURA A1.37
Mapa dos *clusters*



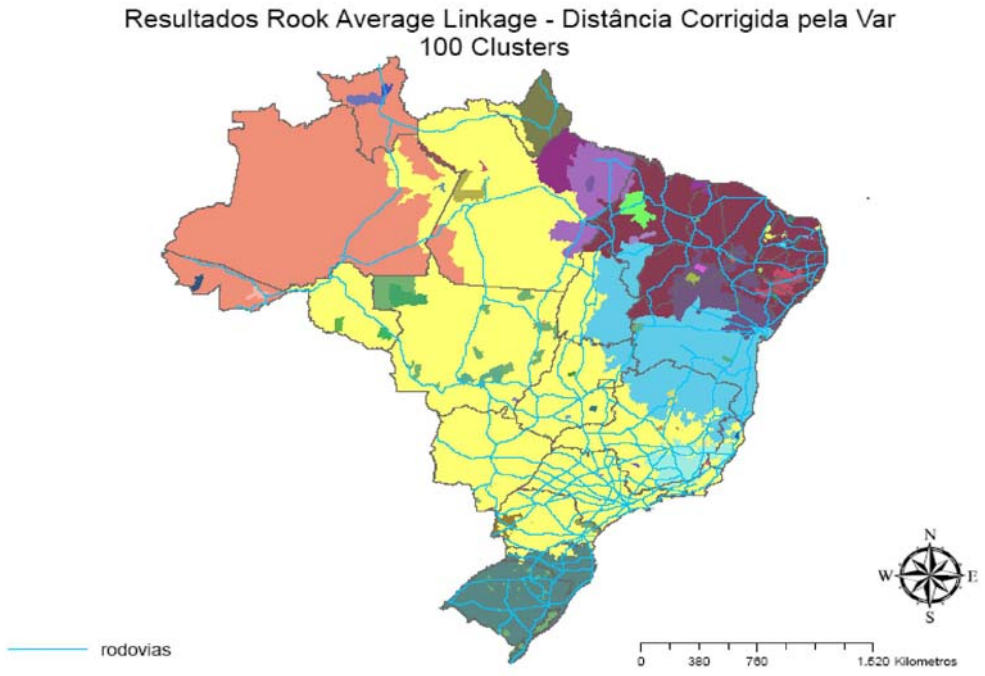
Elaboração dos autores.
Obs.: O mapa está reproduzido conforme o original fornecido pelos autores, cujas características não permitiram melhor ajuste para fim de impressão (nota do Editorial).

FIGURA A1.38
Mapa dos *clusters*



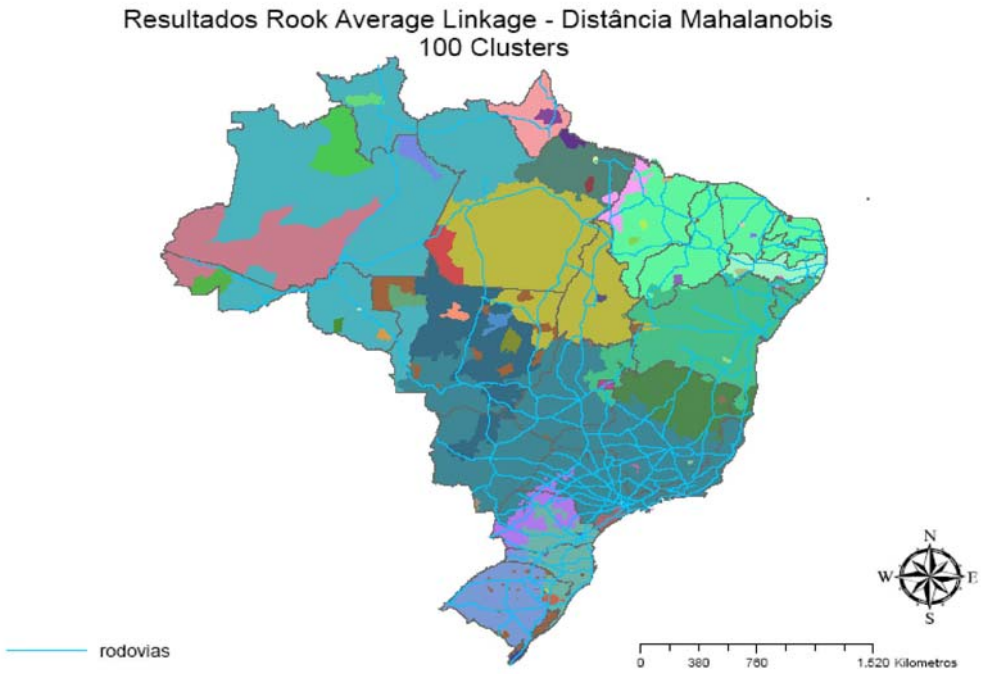
Elaboração dos autores.
Obs.: O mapa está reproduzido conforme o original fornecido pelos autores, cujas características não permitiram melhor ajuste para fim de impressão (nota do Editorial).

FIGURA A1.39
Mapa dos *clusters*



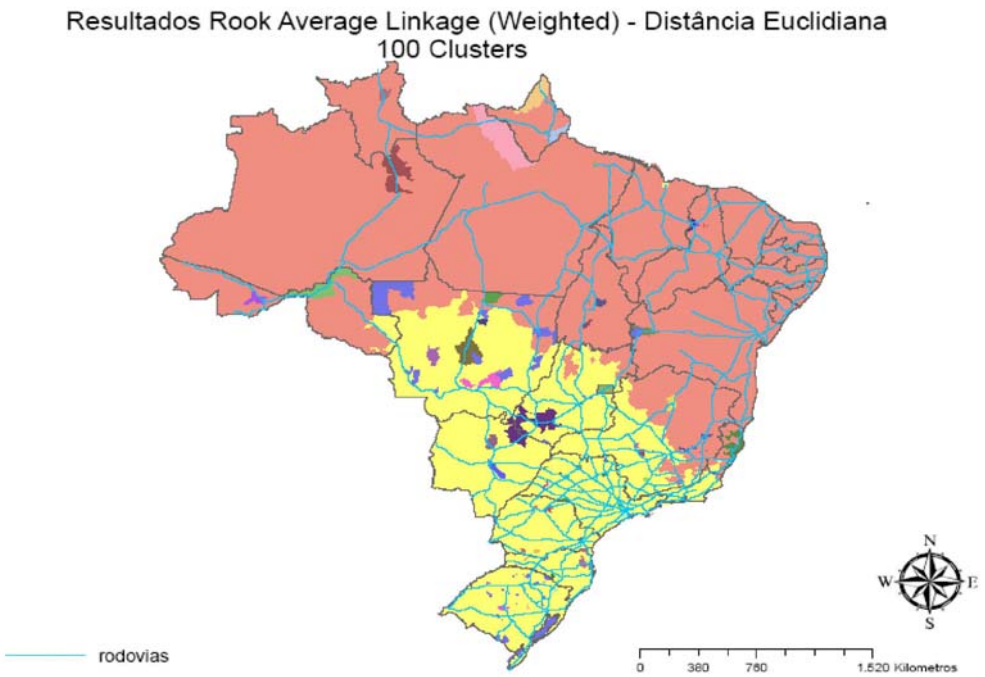
Elaboração dos autores.
Obs.: O mapa está reproduzido conforme o original fornecido pelos autores, cujas características não permitiram melhor ajuste para fim de impressão (nota do Editorial).

FIGURA A1.40
Mapa dos *clusters*



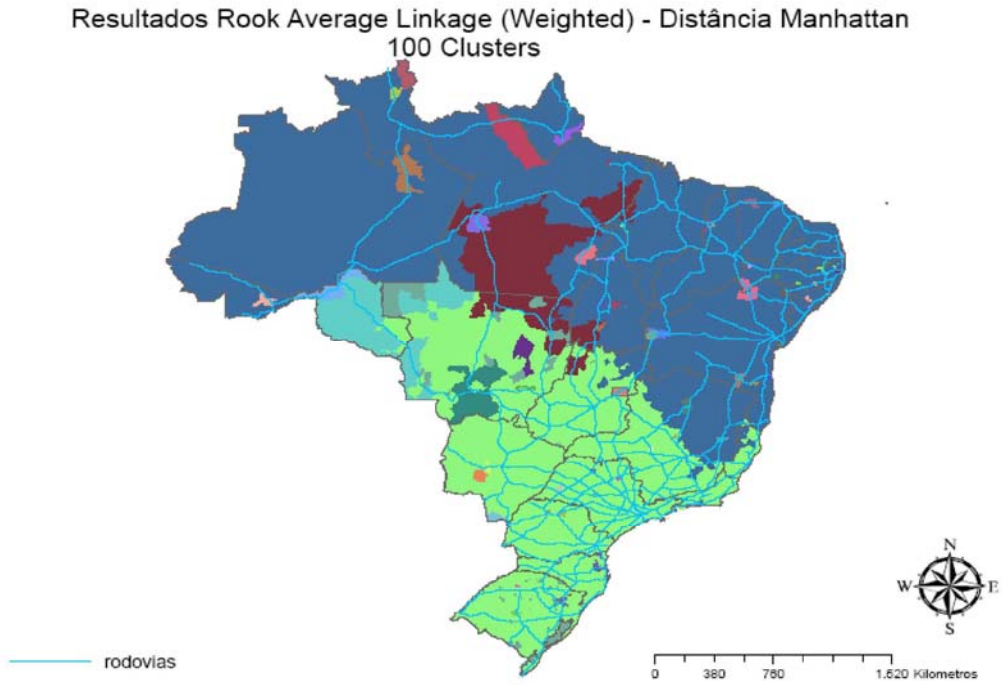
Elaboração dos autores.
Obs.: O mapa está reproduzido conforme o original fornecido pelos autores, cujas características não permitiram melhor ajuste para fim de impressão (nota do Editorial).

FIGURA A1.41
Mapa dos *clusters*



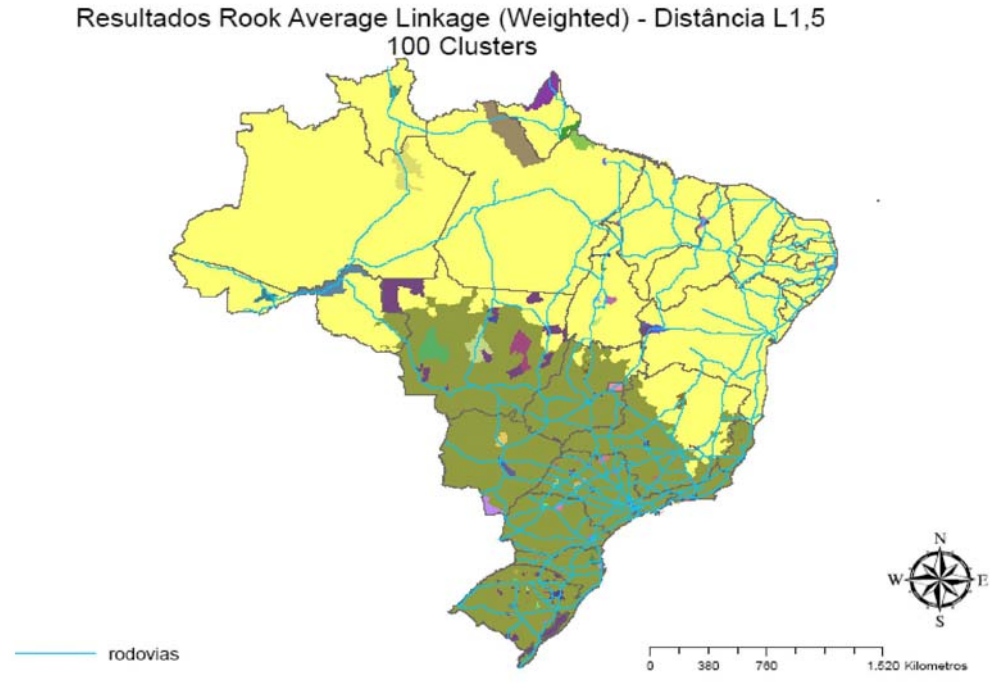
Elaboração dos autores.
Obs.: O mapa está reproduzido conforme o original fornecido pelos autores, cujas características não permitiram melhor ajuste para fim de impressão (nota do Editorial).

FIGURA A1.42
Mapa dos *clusters*



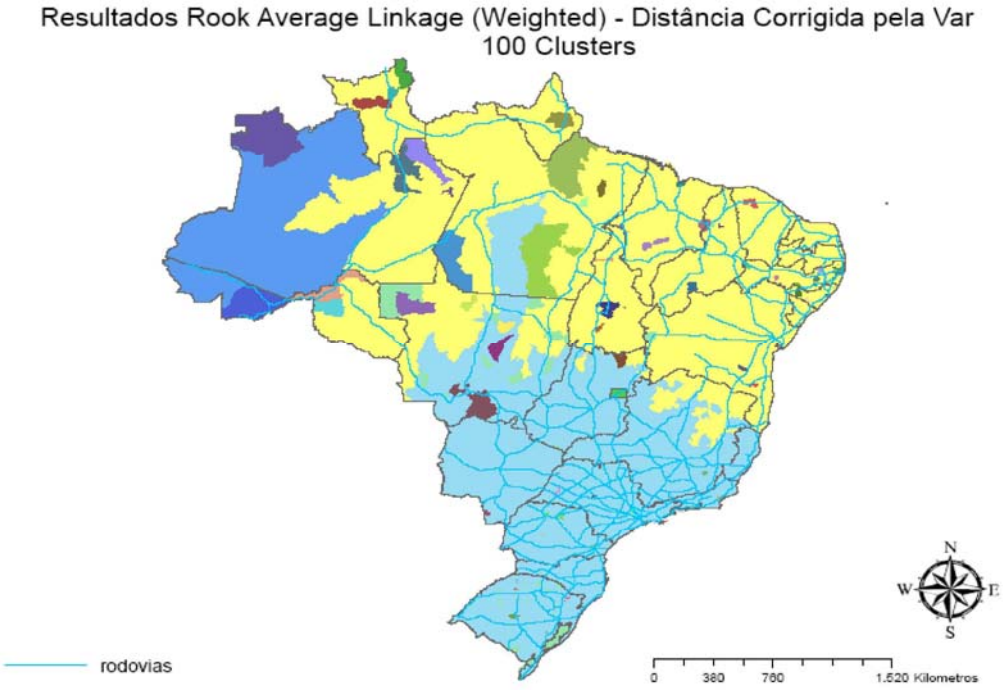
Elaboração dos autores.
Obs.: O mapa está reproduzido conforme o original fornecido pelos autores, cujas características não permitiram melhor ajuste para fim de impressão (nota do Editorial).

FIGURA A1.43
Mapa dos *clusters*



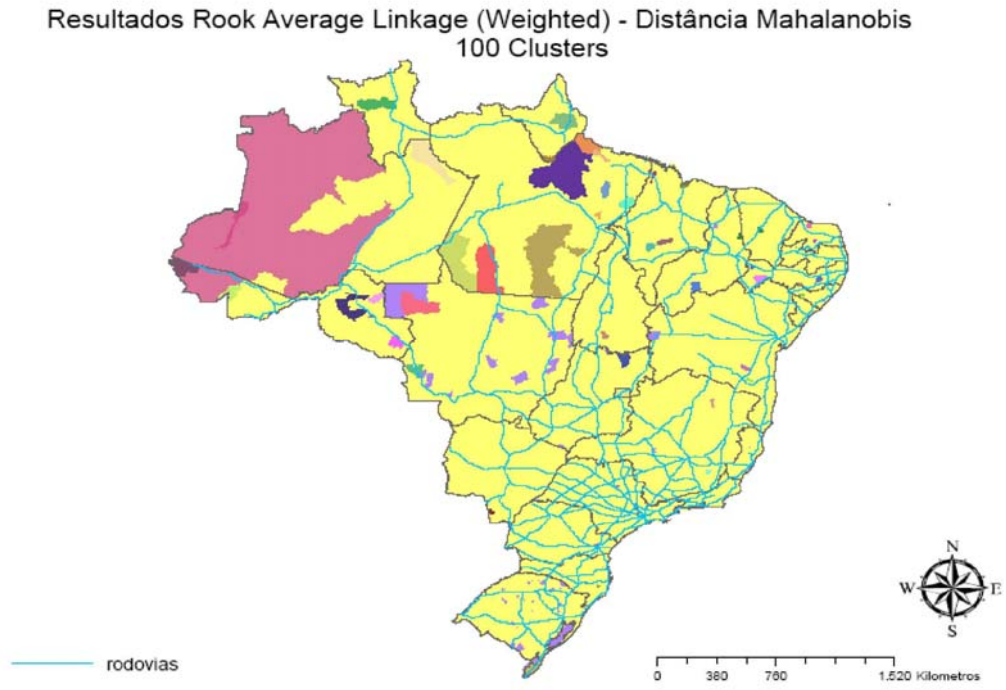
Elaboração dos autores.
Obs.: O mapa está reproduzido conforme o original fornecido pelos autores, cujas características não permitiram melhor ajuste para fim de impressão (nota do Editorial).

FIGURA A1.44
Mapa dos *clusters*



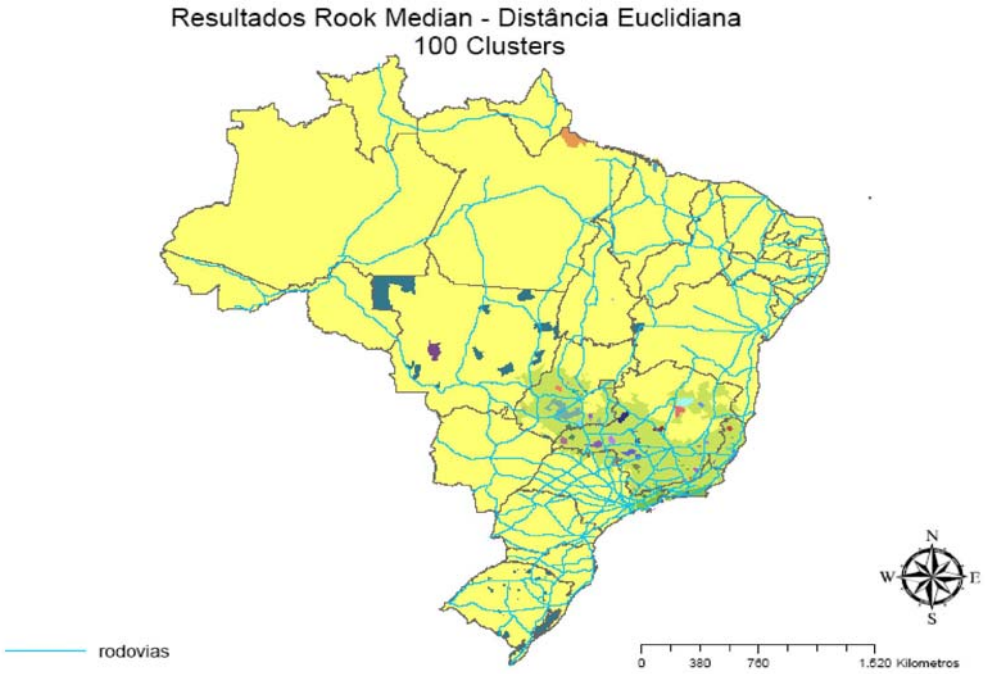
Elaboração dos autores.
Obs.: O mapa está reproduzido conforme o original fornecido pelos autores, cujas características não permitiram melhor ajuste para fim de impressão (nota do Editorial).

FIGURA A1.45
Mapa dos *clusters*



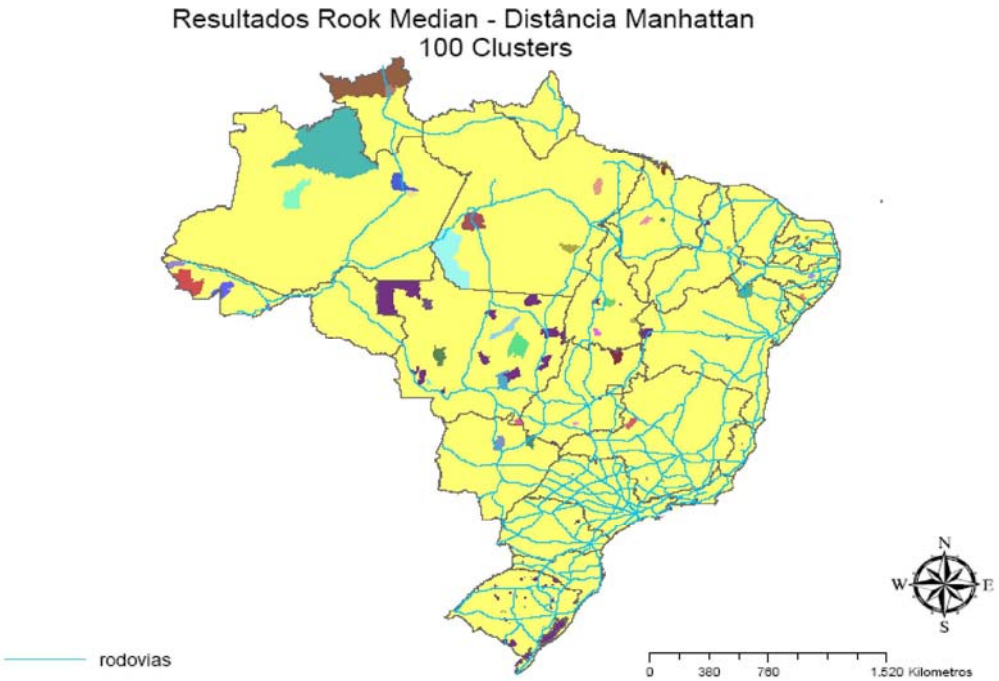
Elaboração dos autores.
Obs.: O mapa está reproduzido conforme o original fornecido pelos autores, cujas características não permitiram melhor ajuste para fim de impressão (nota do Editorial).

FIGURA A1.46
Mapa dos *clusters*



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).

FIGURA A1.47
Mapa dos *clusters*

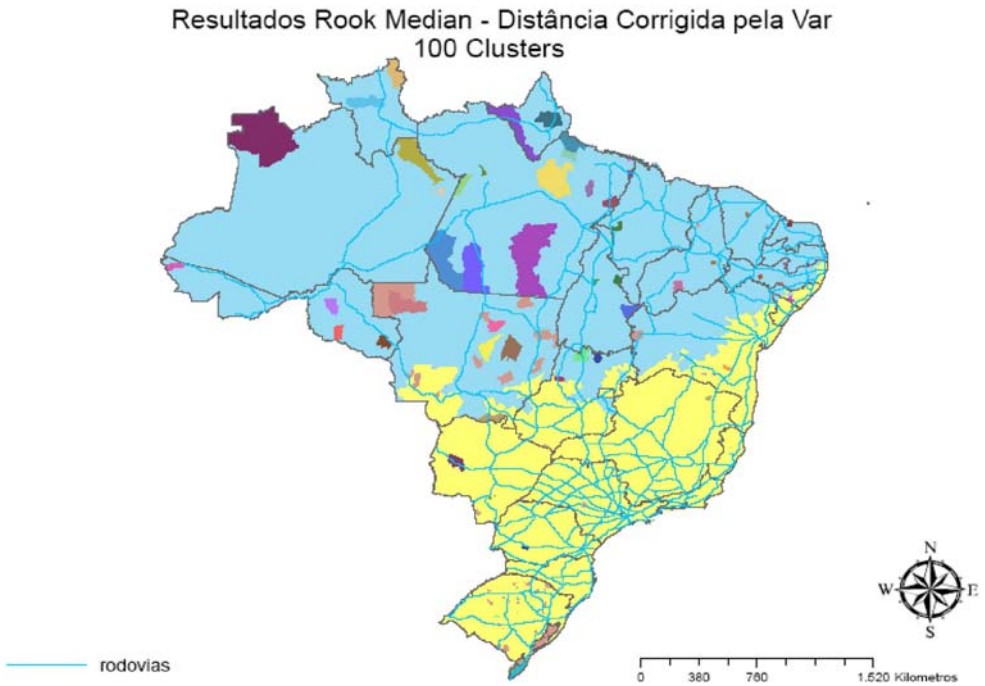


Elaboração dos autores.
Obs.: O mapa está reproduzido conforme o original fornecido pelos autores, cujas características não permitiram melhor ajuste para fim de impressão (nota do Editorial).

FIGURA A1.48
Mapa dos *clusters*

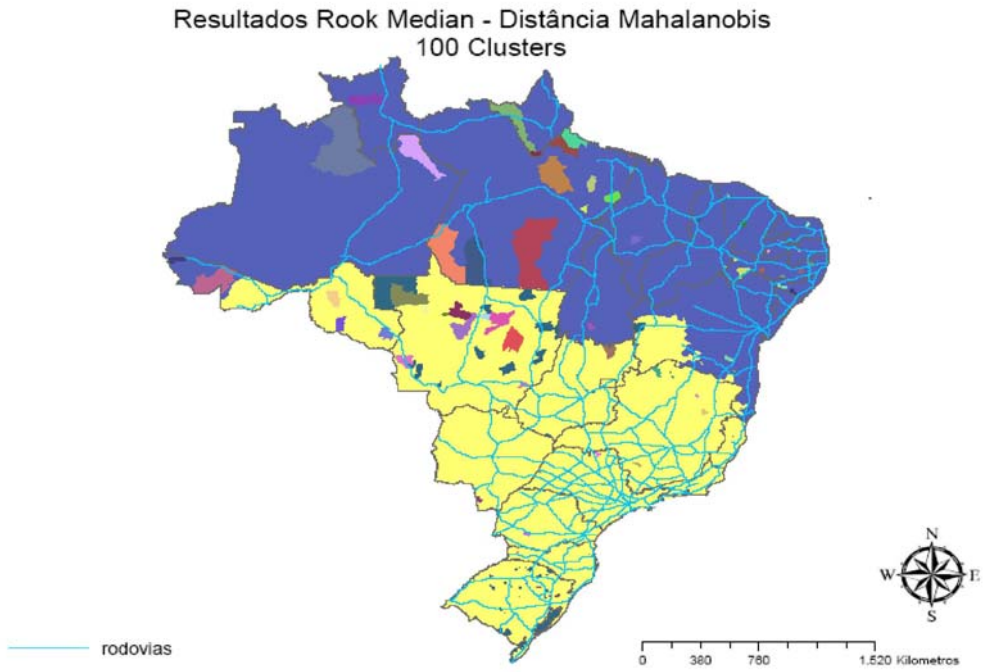
Elaboração dos autores.
Obs.: O mapa está reproduzido conforme o original fornecido pelos autores, cujas características não permitiram melhor ajuste para fim de impressão (nota do Editorial).

FIGURA A1.49
Mapa dos *clusters*



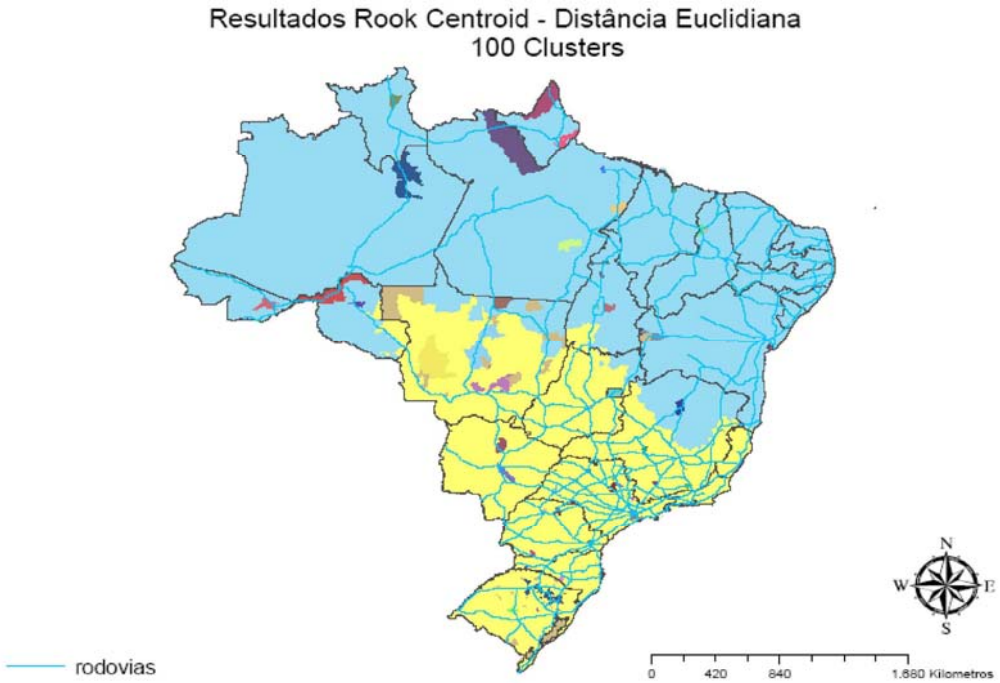
Elaboração dos autores.
Obs.: O mapa está reproduzido conforme o original fornecido pelos autores, cujas características não permitiram melhor ajuste para fim de impressão (nota do Editorial).

FIGURA A1.50
Mapa dos *clusters*



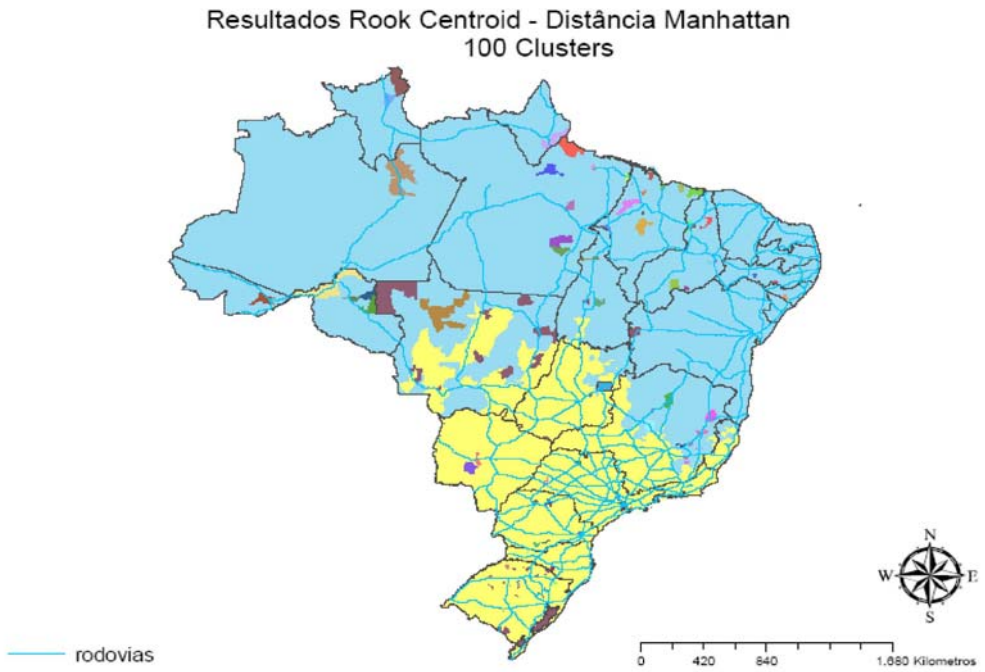
Elaboração dos autores.
Obs.: O mapa está reproduzido conforme o original fornecido pelos autores, cujas características não permitiram melhor ajuste para fim de impressão (nota do Editorial).

FIGURA A1.51
Mapa dos *clusters*



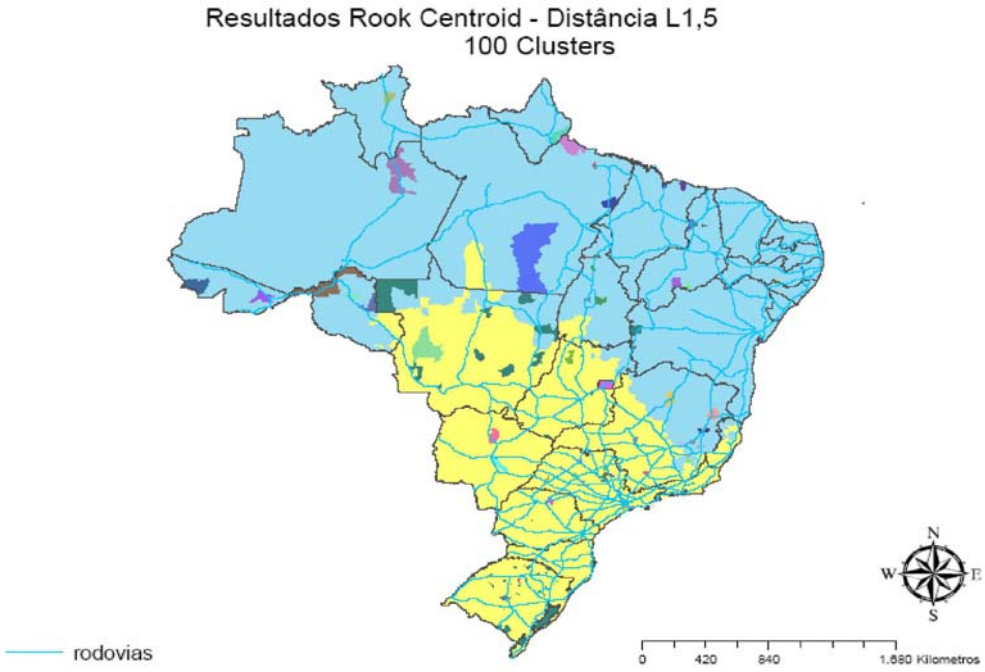
Elaboração dos autores.
Obs.: O mapa está reproduzido conforme o original fornecido pelos autores, cujas características não permitiram melhor ajuste para fim de impressão (nota do Editorial).

FIGURA A1.52
Mapa dos *clusters*



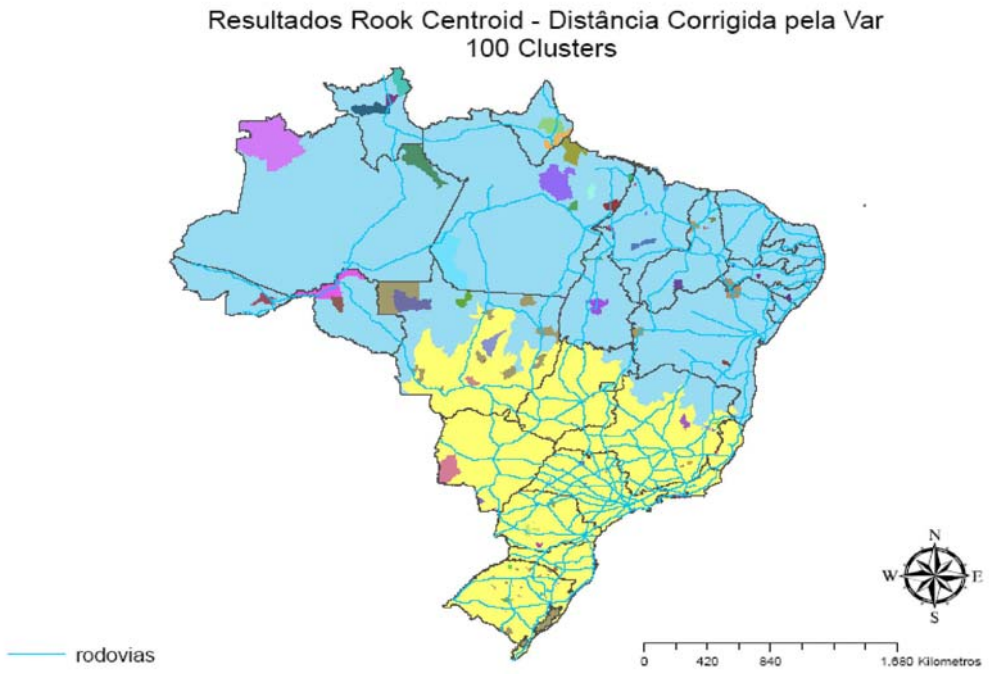
Elaboração dos autores.
Obs.: O mapa está reproduzido conforme o original fornecido pelos autores, cujas características não permitiram melhor ajuste para fim de impressão (nota do Editorial).

FIGURA A1.53
Mapa dos *clusters*



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).

FIGURA A1.54
Mapa dos *clusters*

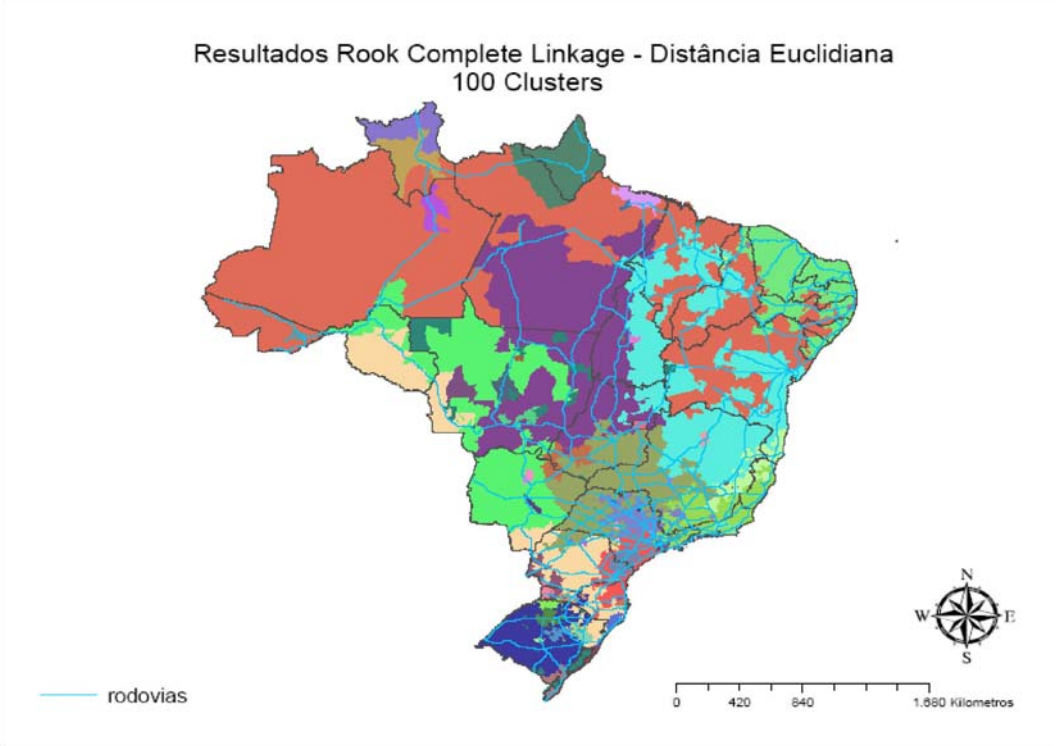


Elaboração dos autores.
Obs.: O mapa está reproduzido conforme o original fornecido pelos autores, cujas características não permitiram melhor ajuste para fim de impressão (nota do Editorial).

FIGURA A1.55
Mapa dos *clusters*

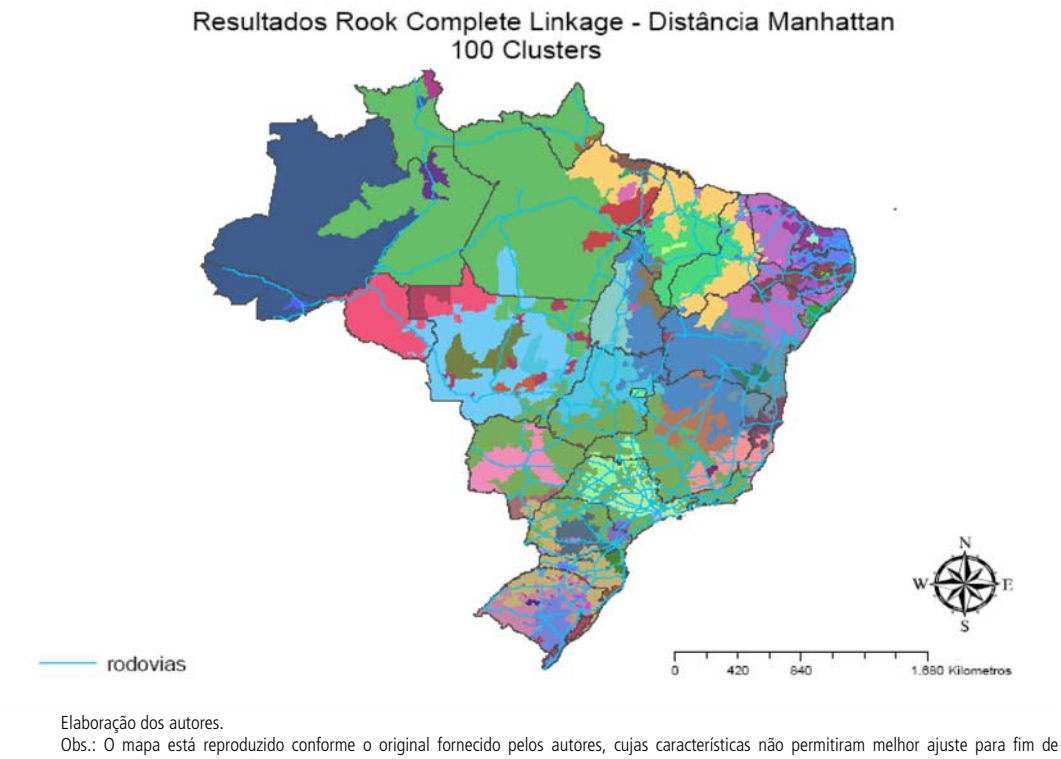
Elaboração dos autores.
Obs.: O mapa está reproduzido conforme o original fornecido pelos autores, cujas características não permitiram melhor ajuste para fim de impressão (nota do Editorial).

FIGURA A1.56
Mapa dos *clusters*



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).

FIGURA A1.57
Mapa dos *clusters*

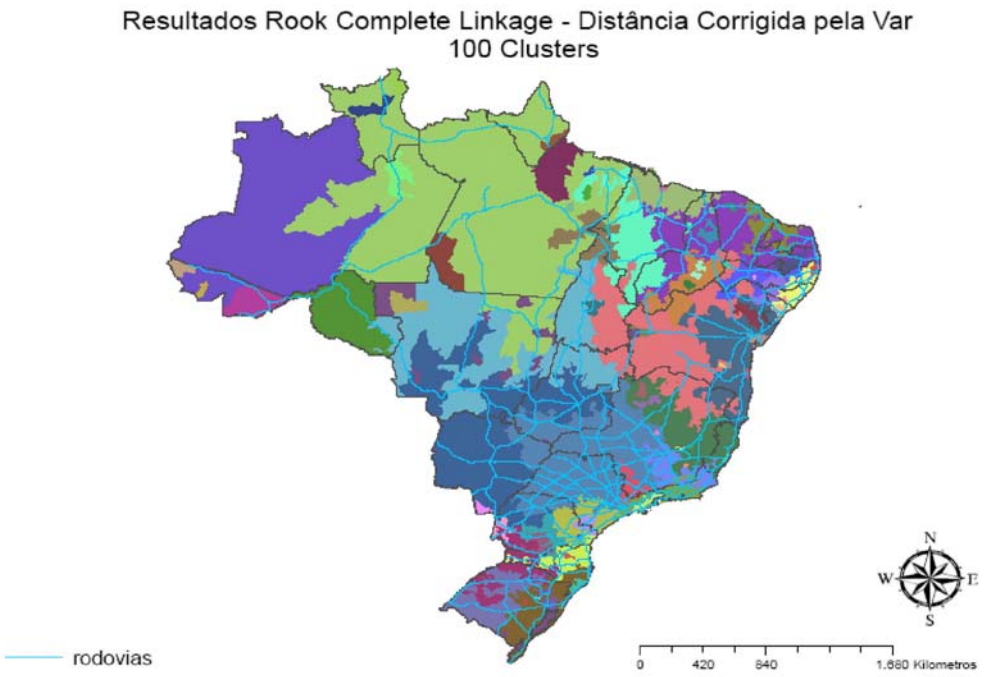


Elaboração dos autores.
Obs.: O mapa está reproduzido conforme o original fornecido pelos autores, cujas características não permitiram melhor ajuste para fim de impressão (nota do Editorial).

FIGURA A1.58
Mapa dos *clusters*

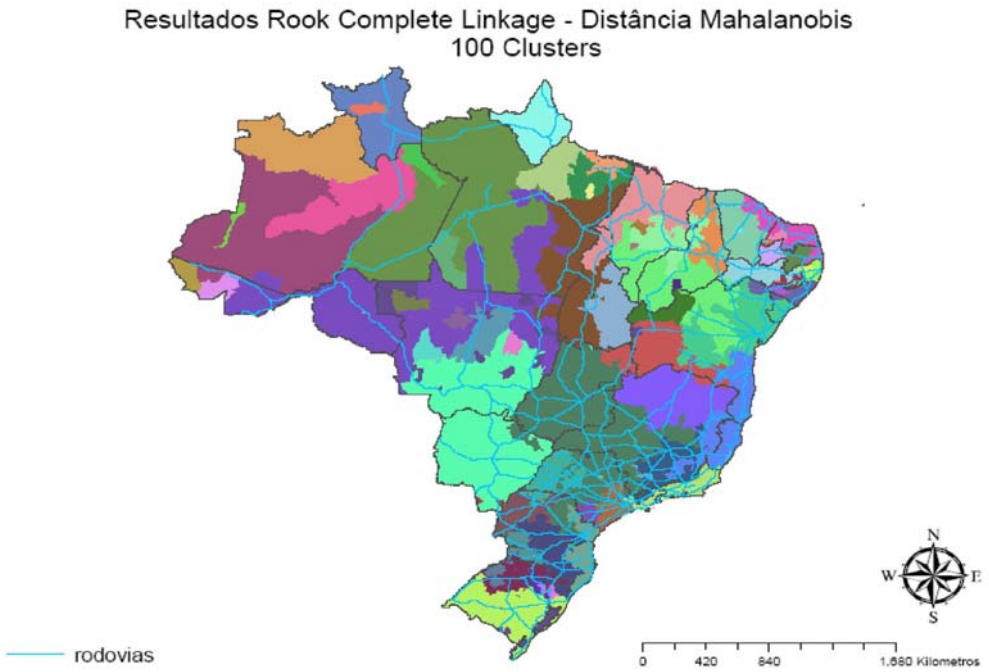
Elaboração dos autores.
Obs.: O mapa está reproduzido conforme o original fornecido pelos autores, cujas características não permitiram melhor ajuste para fim de impressão (nota do Editorial).

FIGURA A1.59
Mapa dos *clusters*



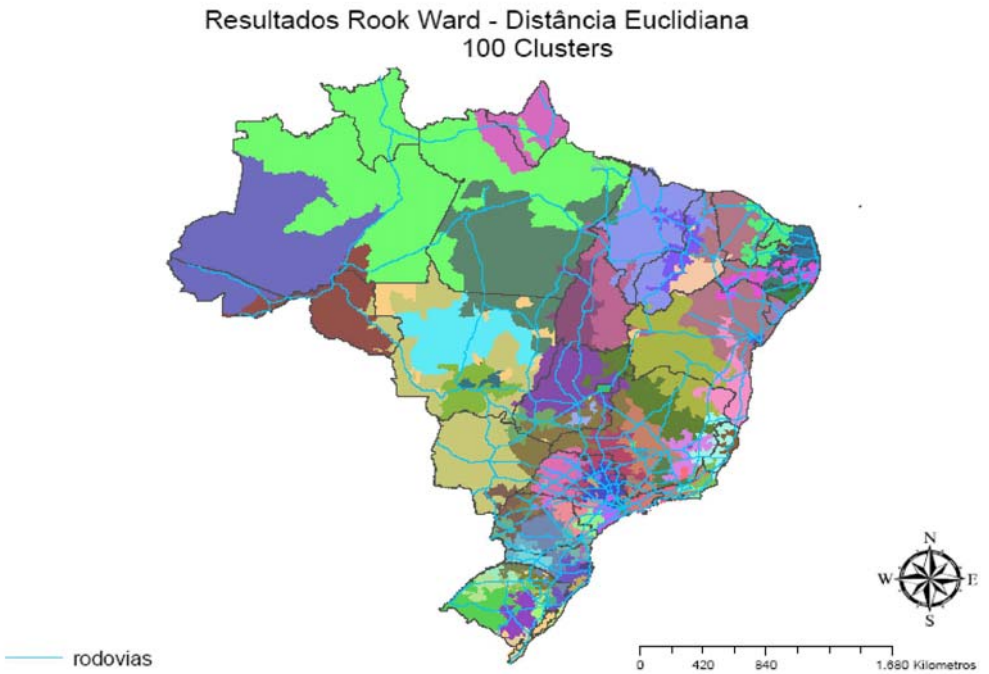
Elaboração dos autores.
Obs.: O mapa está reproduzido conforme o original fornecido pelos autores, cujas características não permitiram melhor ajuste para fim de impressão (nota do Editorial).

FIGURA A1.60
Mapa dos *clusters*



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).

FIGURA A1.61
Mapa dos clusters



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).

FIGURA A1.62
Mapa dos *clusters*

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).

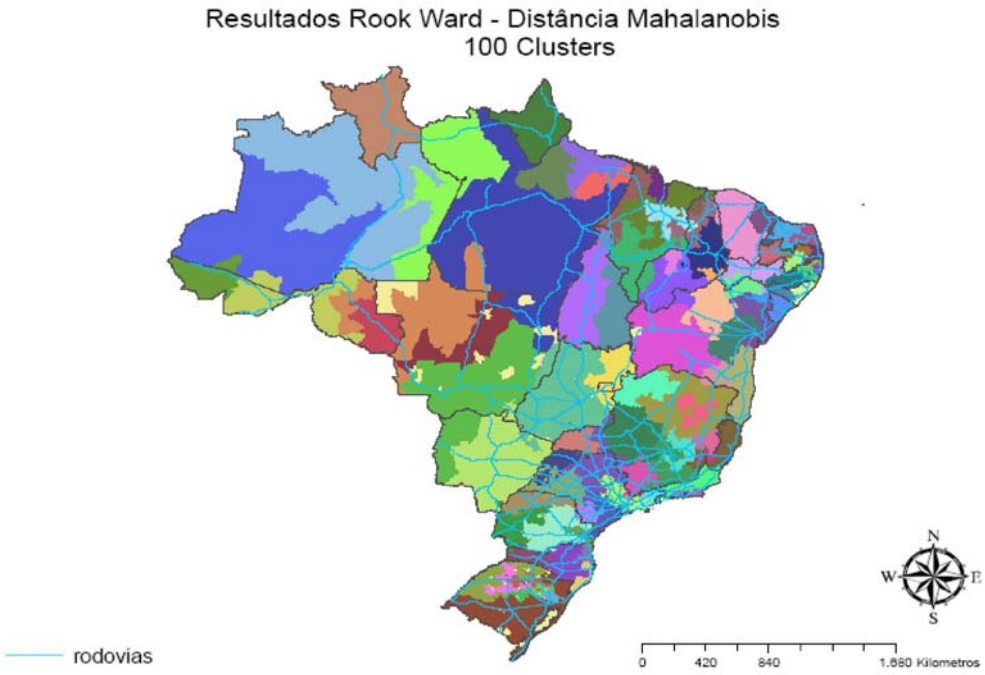
FIGURA A1.63
Mapa dos *clusters*

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).

FIGURA A1.64
Mapa dos *clusters*

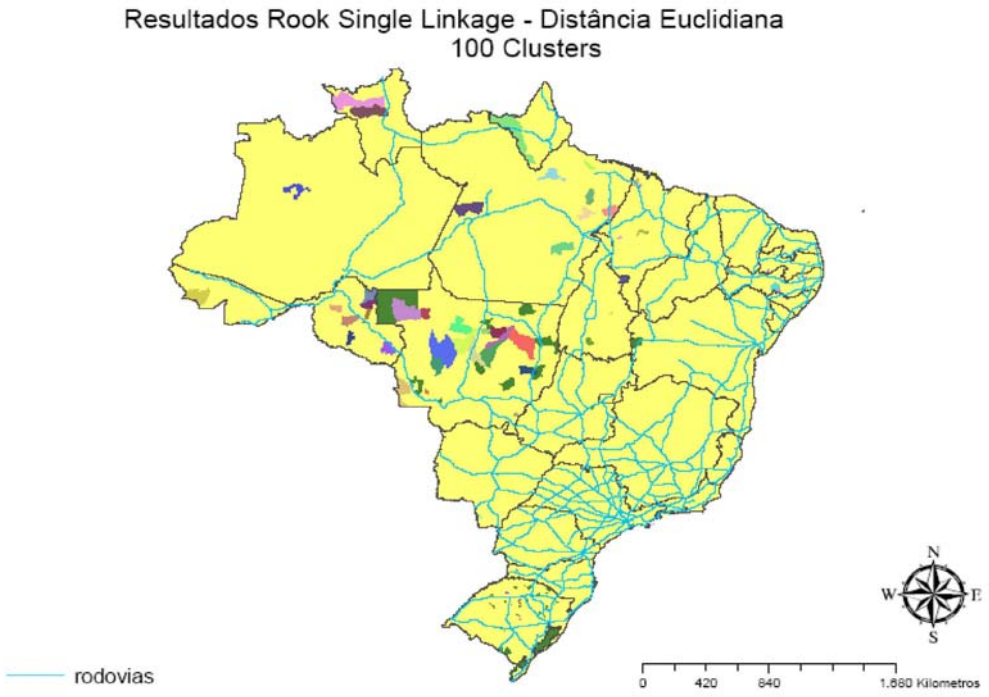
Elaboração dos autores.
Obs.: O mapa está reproduzido conforme o original fornecido pelos autores, cujas características não permitiram melhor ajuste para fim de impressão (nota do Editorial).

FIGURA A1.65
Mapa dos *clusters*



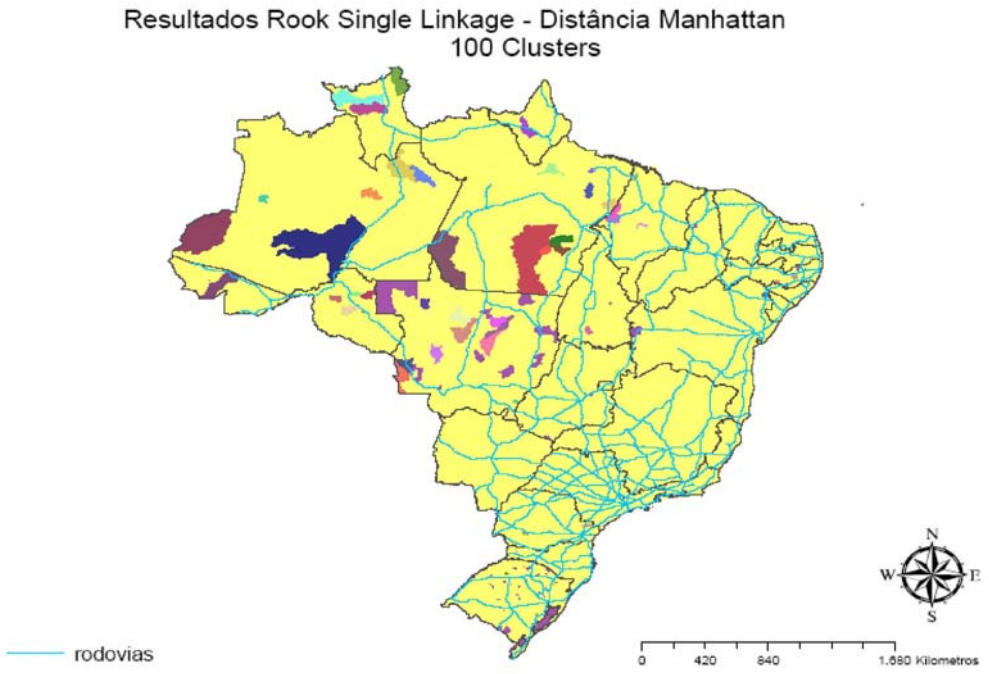
Elaboração dos autores.
Obs.: O mapa está reproduzido conforme o original fornecido pelos autores, cujas características não permitiram melhor ajuste para fim de impressão (nota do Editorial).

FIGURA A1.66
Mapa dos *clusters*



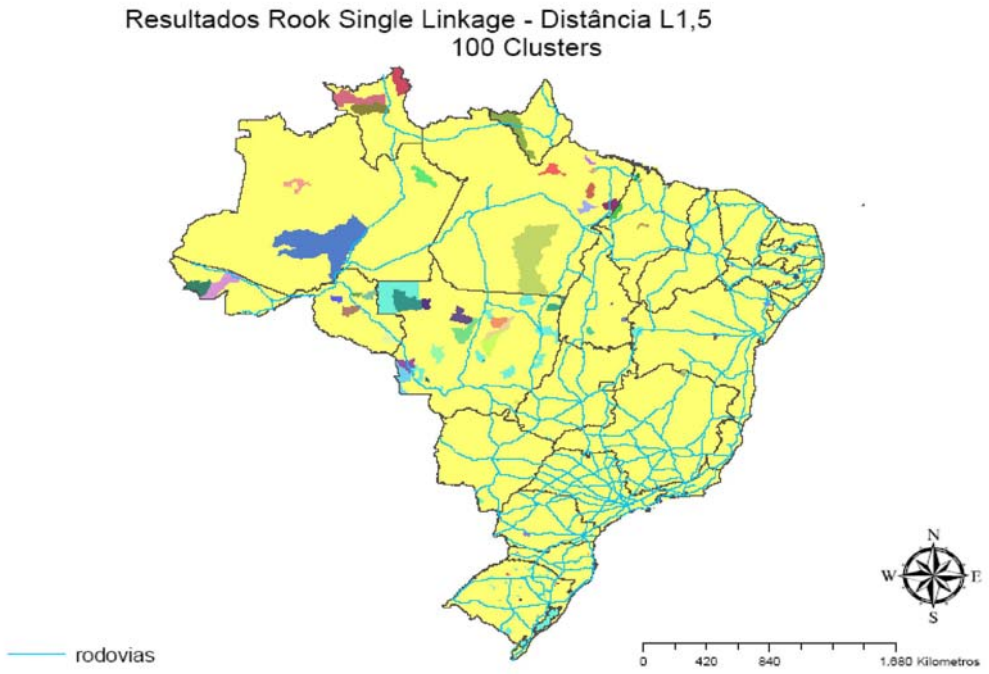
Elaboração dos autores.
Obs.: O mapa está reproduzido conforme o original fornecido pelos autores, cujas características não permitiram melhor ajuste para fim de impressão (nota do Editorial).

FIGURA A1.67
Mapa dos *clusters*



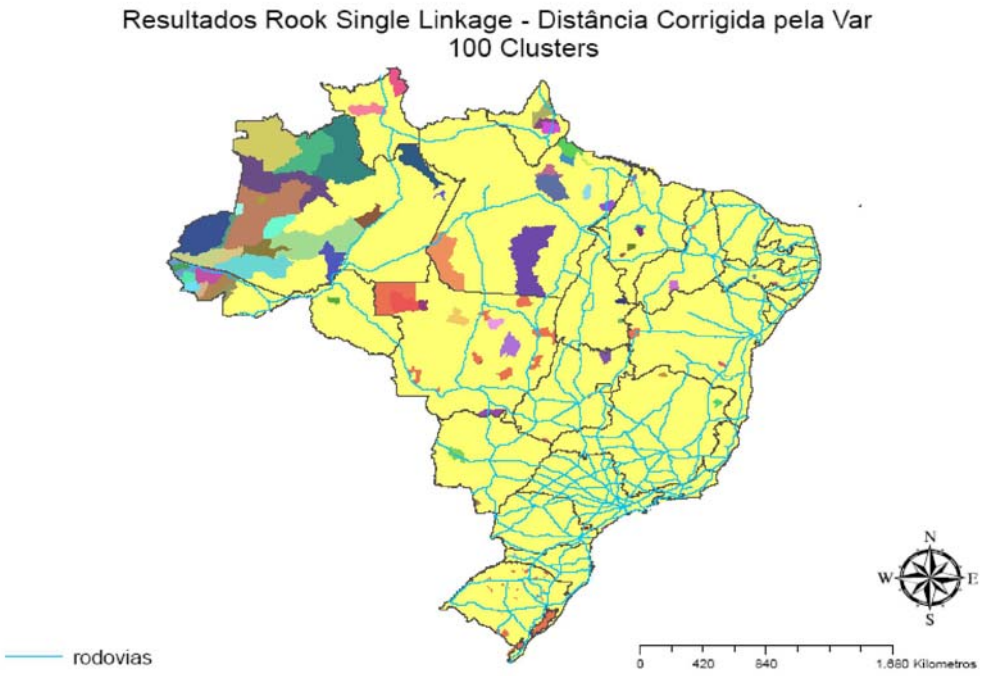
Elaboração dos autores.
Obs.: O mapa está reproduzido conforme o original fornecido pelos autores, cujas características não permitiram melhor ajuste para fim de impressão (nota do Editorial).

FIGURA A1.68
Mapa dos *clusters*



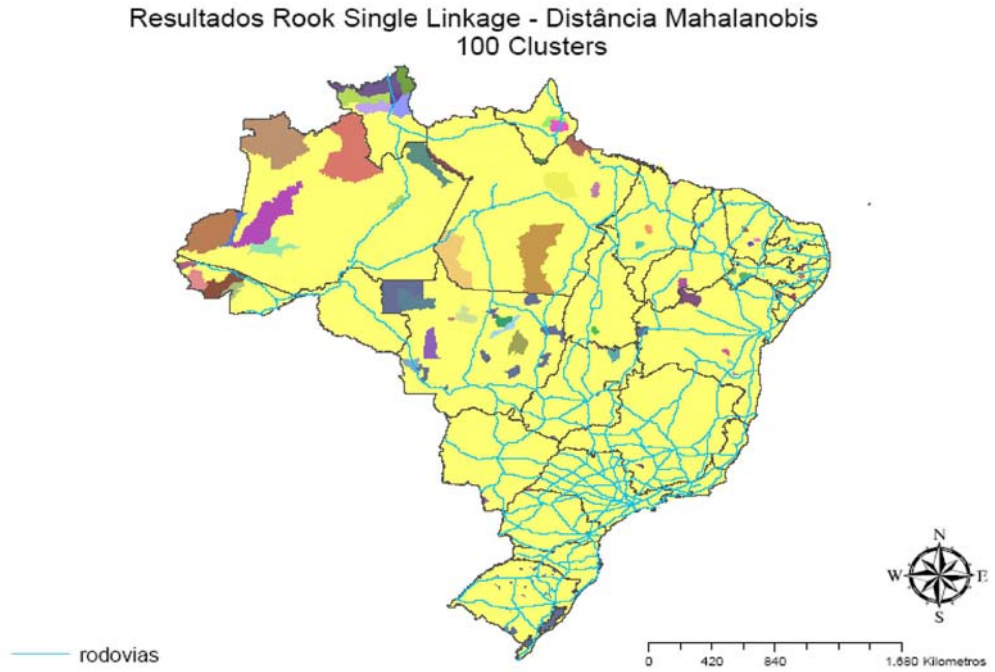
Elaboração dos autores.
Obs.: O mapa está reproduzido conforme o original fornecido pelos autores, cujas características não permitiram melhor ajuste para fim de impressão (nota do Editorial).

FIGURA A1.69
Mapa dos *clusters*



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).

FIGURA A1.70
Mapa dos *clusters*



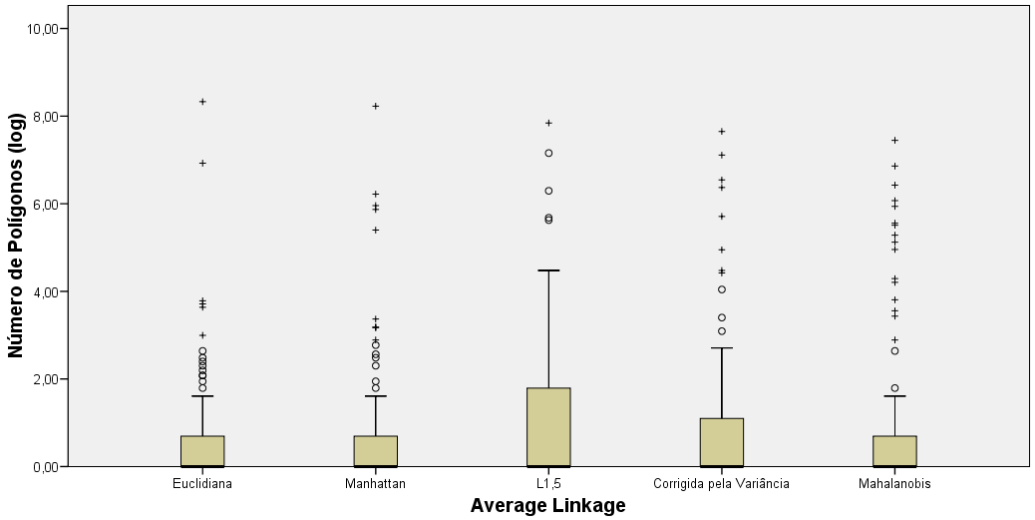
Elaboração dos autores.
Obs.: O mapa está reproduzido conforme o original fornecido pelos autores, cujas características não permitiram melhor ajuste para fim de impressão (nota do Editorial).

ANEXO 2

BOX PLOTS PARA AVALIAÇÃO DO NÚMERO DE MUNICÍPIOS POR *CLUSTER*[3]

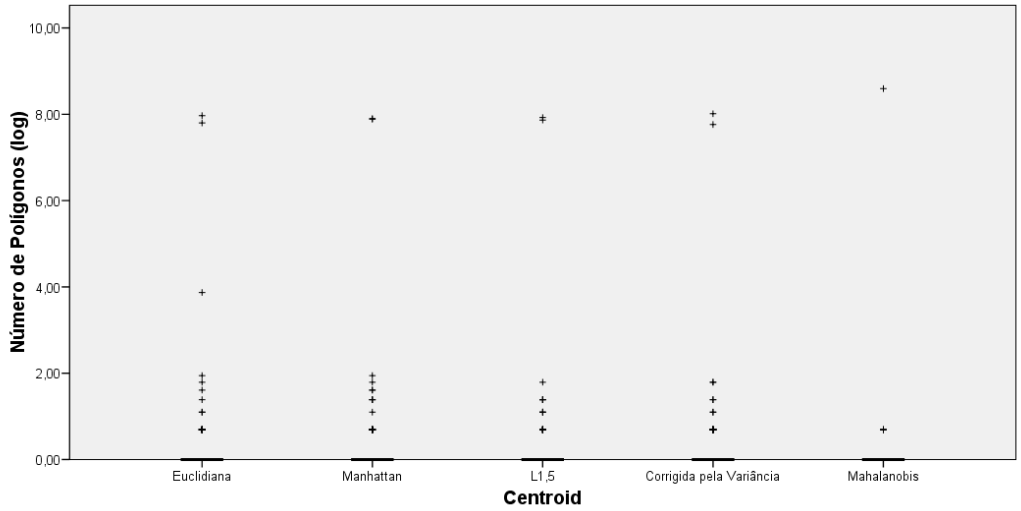
Vizinhança do tipo *rook*

FIGURA A2.1
Distribuição do número de municípios – método *average linkage*



Elaboração dos autores.

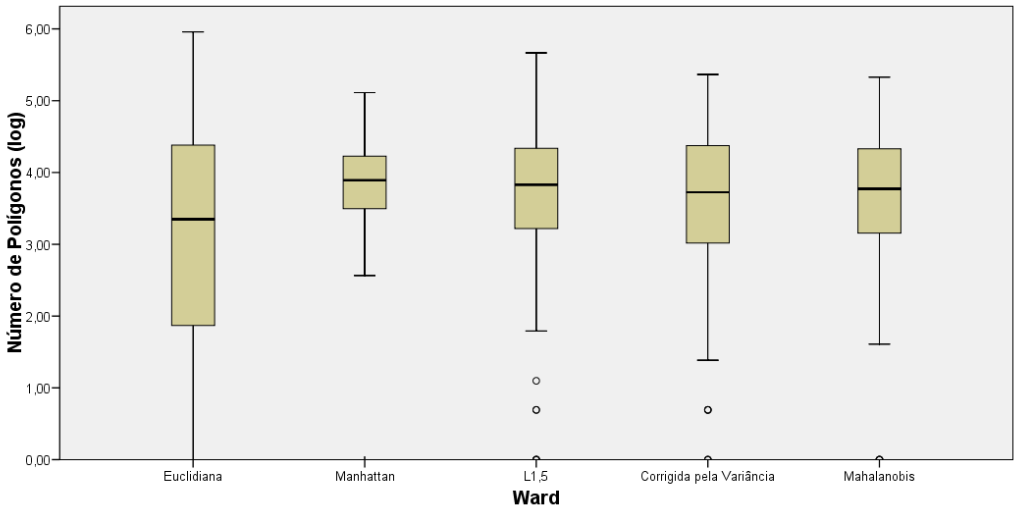
FIGURA A2.2
Distribuição do número de municípios – método *centroid*



Elaboração dos autores.

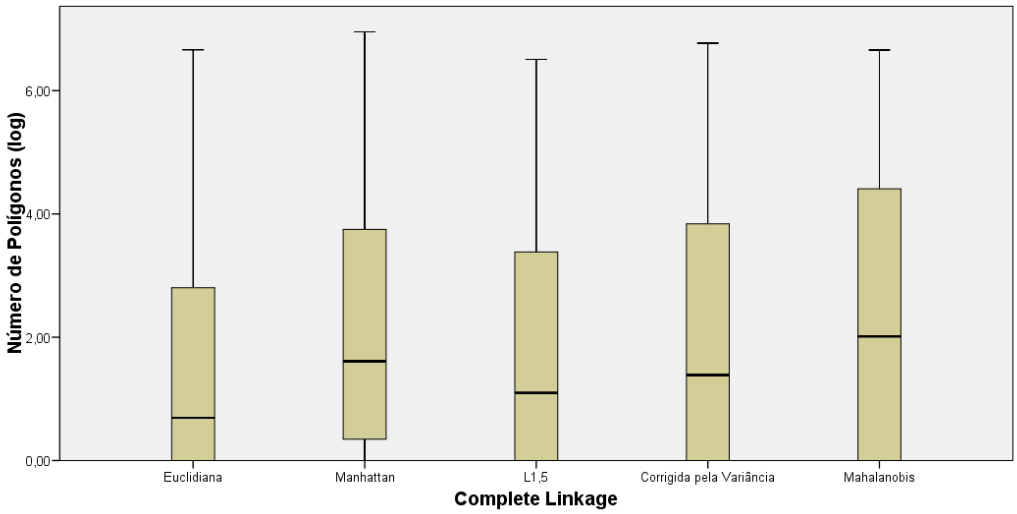
3. Para visualização em cores, acessar a seção *O trabalho do Ipea*, subseção *Publicações*, no site: <http://www.ipea.gov.br>.

FIGURA A2.3
Distribuição do número de municípios – método de Ward



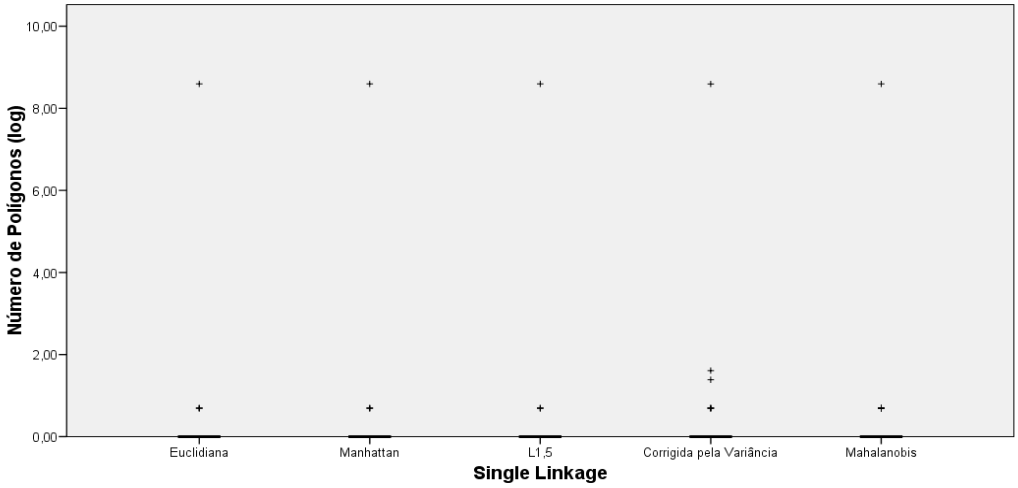
Elaboração dos autores.

FIGURA A2.4
Distribuição do número de municípios – método *complete linkage*



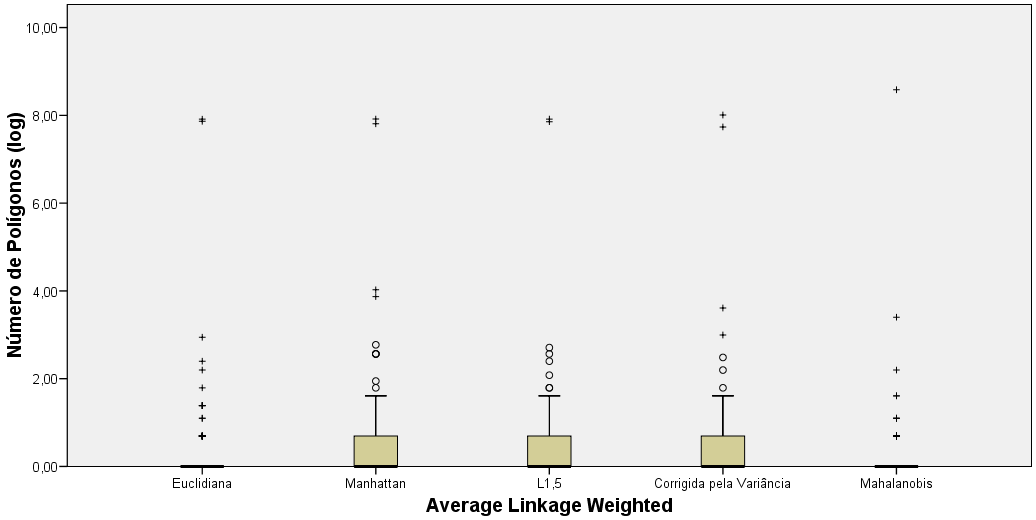
Elaboração dos autores.

FIGURA A2.5
Distribuição do número de municípios – método *single linkage*



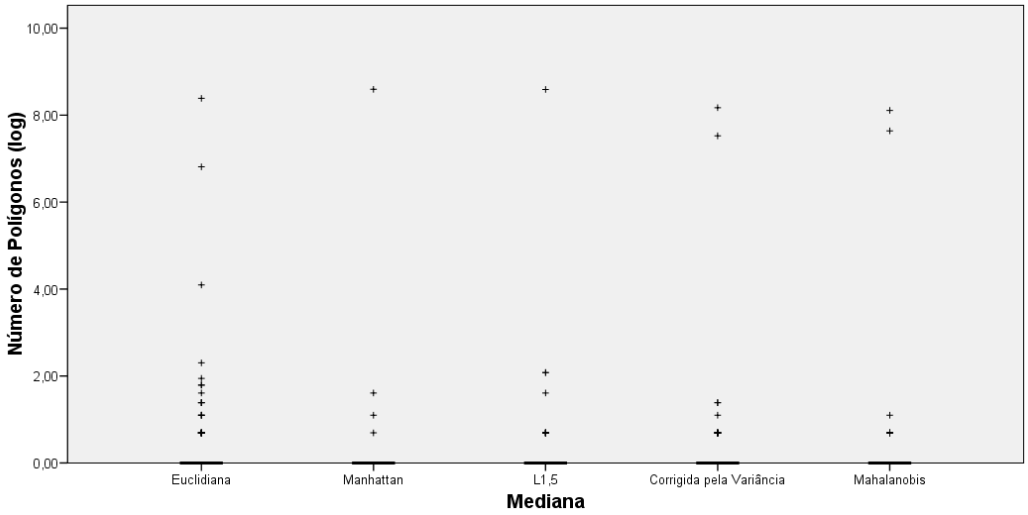
Elaboração dos autores.

FIGURA A2.6
Distribuição do número de municípios – método *average linkage weighted*



Elaboração dos autores.

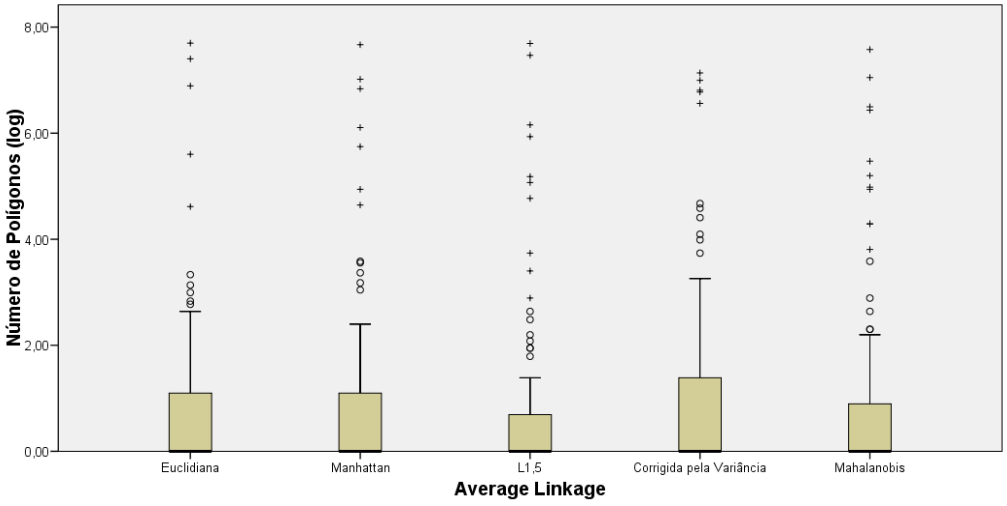
FIGURA A2.7
Distribuição do número de municípios – método da mediana



Elaboração dos autores.

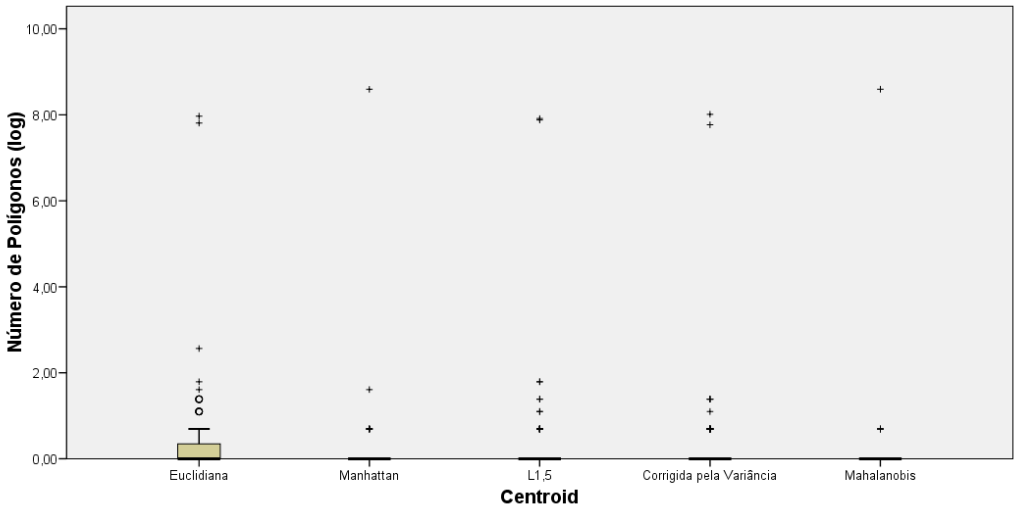
Vizinhança do tipo *queen*

FIGURA A2.8
Distribuição do número de municípios – método *average linkage*



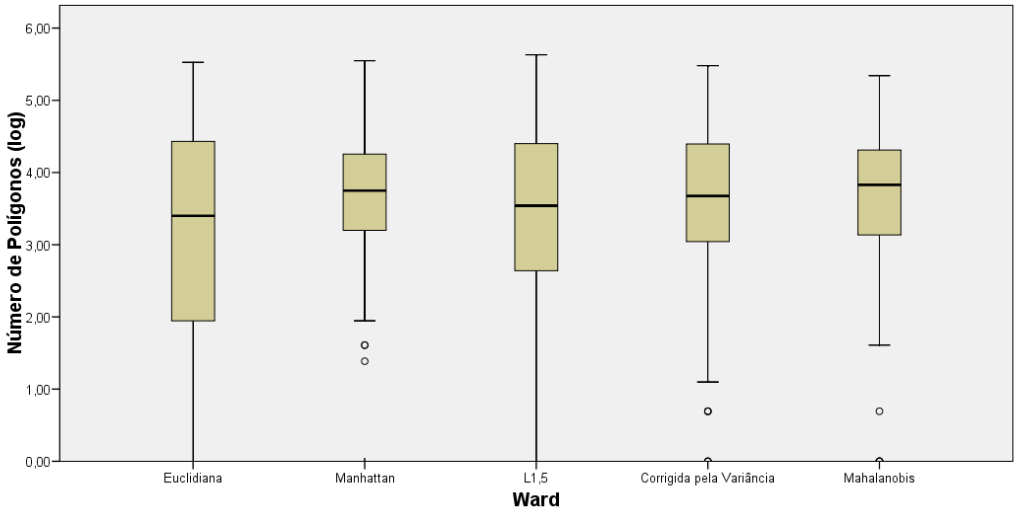
Elaboração dos autores.

FIGURA A2.9
Distribuição do número de municípios – método *centroid*



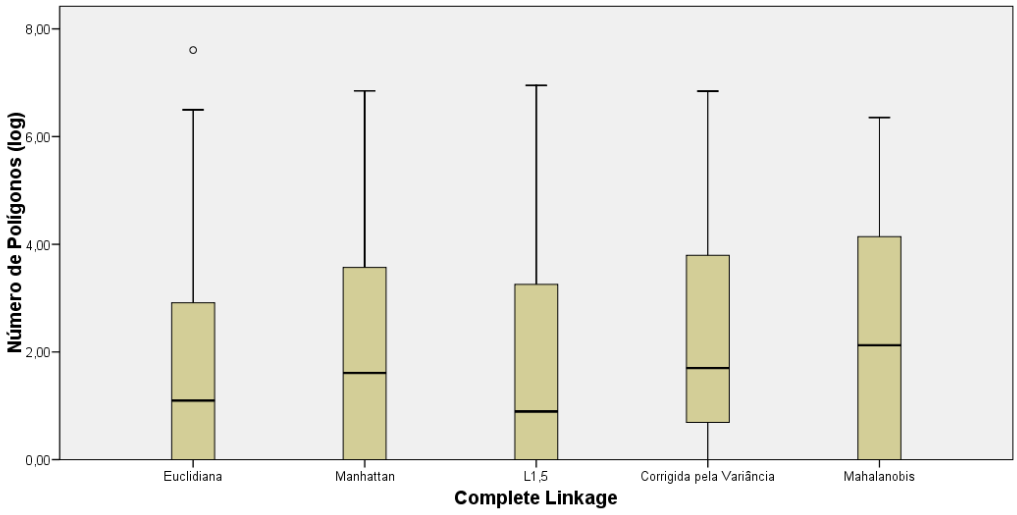
Elaboração dos autores.

FIGURA A2.10
Distribuição do número de municípios – método de Ward



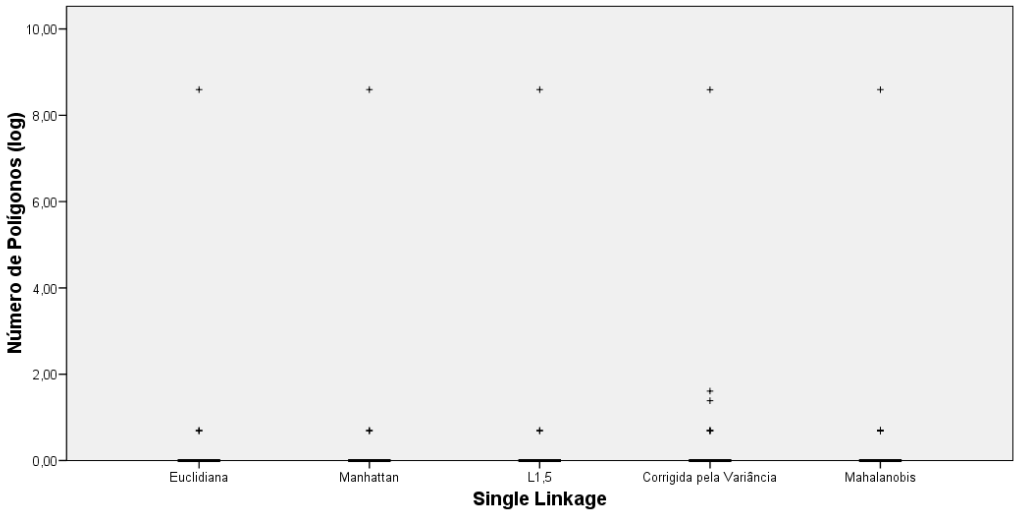
Elaboração dos autores.

FIGURA A2.11
Distribuição do número de municípios – método *complete linkage*



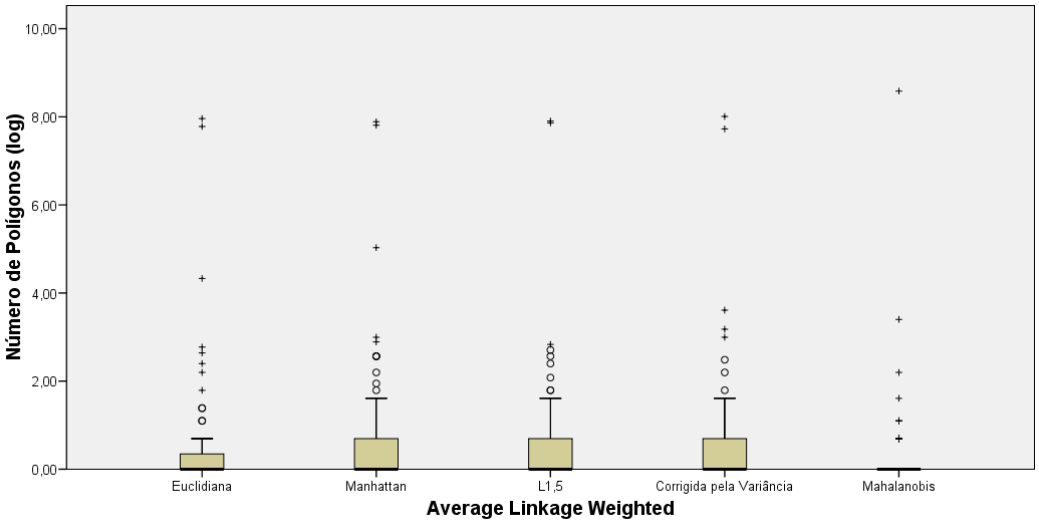
Elaboração dos autores.

FIGURA A2.12
Distribuição do número de municípios – método *single linkage*



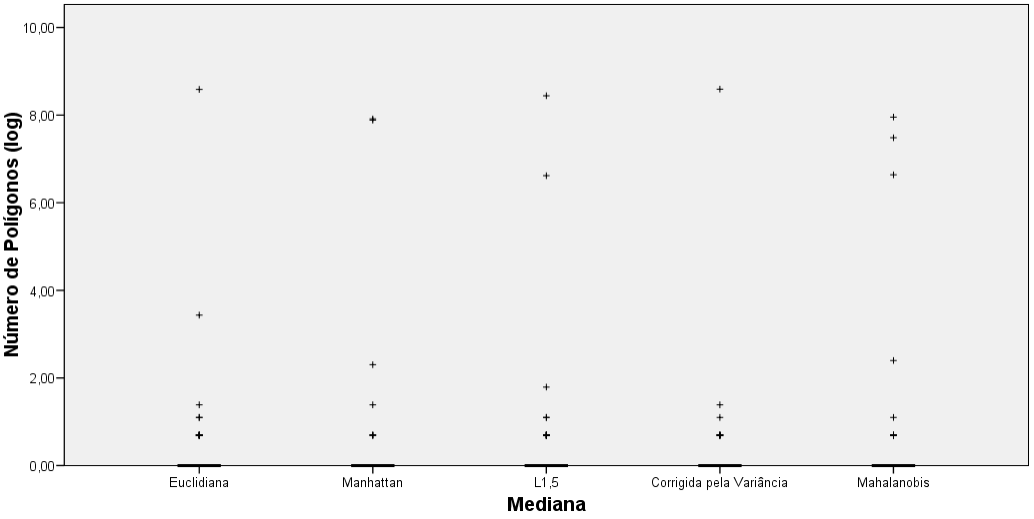
Elaboração dos autores.

FIGURA A2.13
Distribuição do número de municípios – método *average linkage weighted*



Elaboração dos autores.

FIGURA A2.14
Distribuição do número de municípios – método da mediana

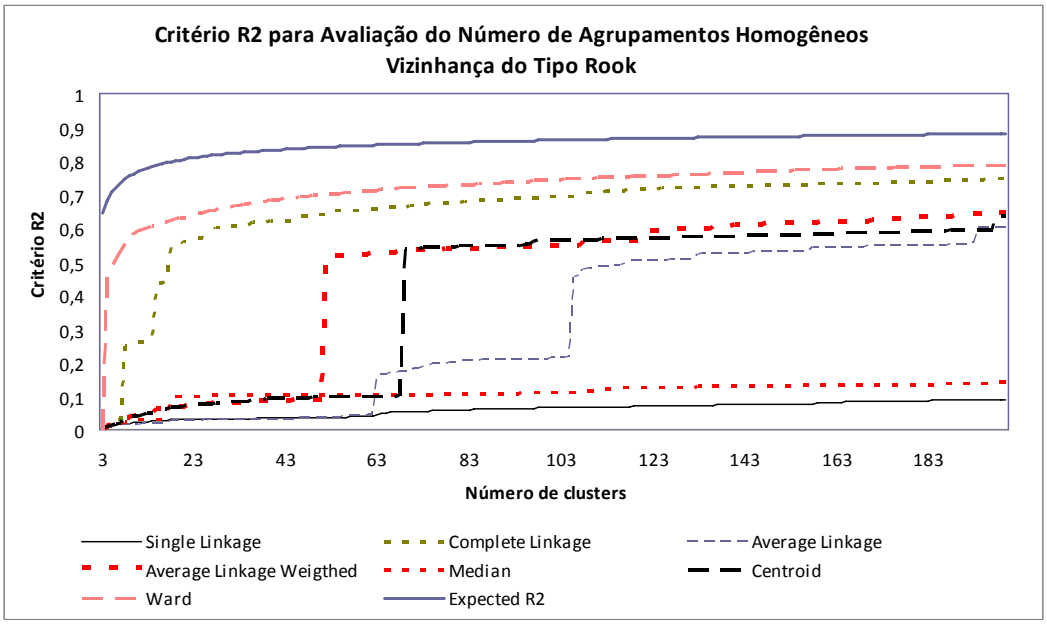


Elaboração dos autores.

ANEXO 3

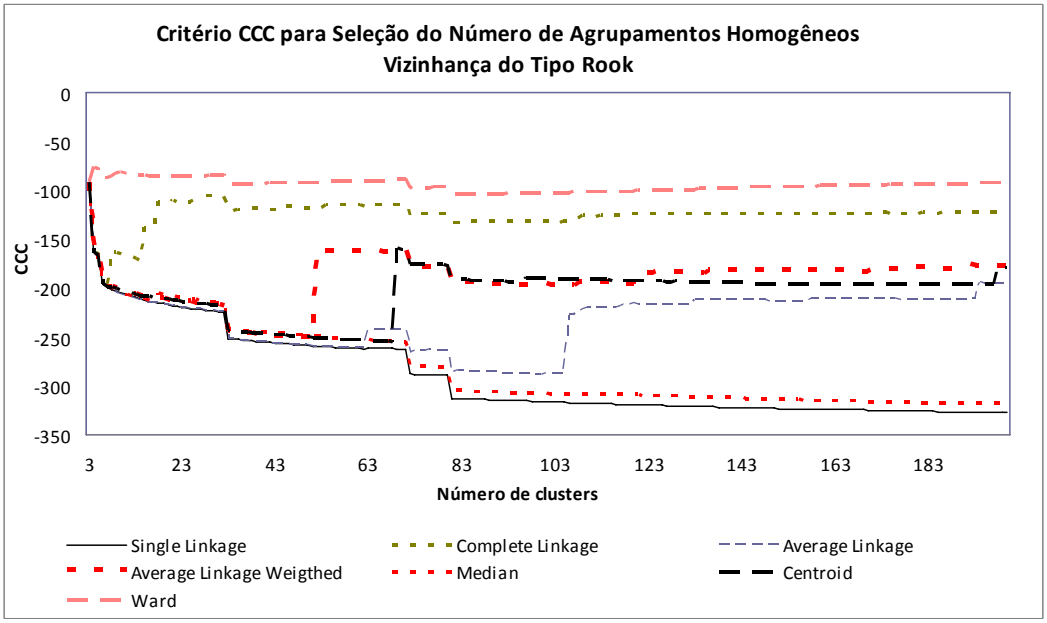
GRÁFICOS DOS CRITÉRIOS DE SELEÇÃO DO NÚMERO DE AGRUPAMENTOS[4]

FIGURA A3.1
Critério *CCC* para vizinhança do tipo *rook*



Elaboração dos autores.

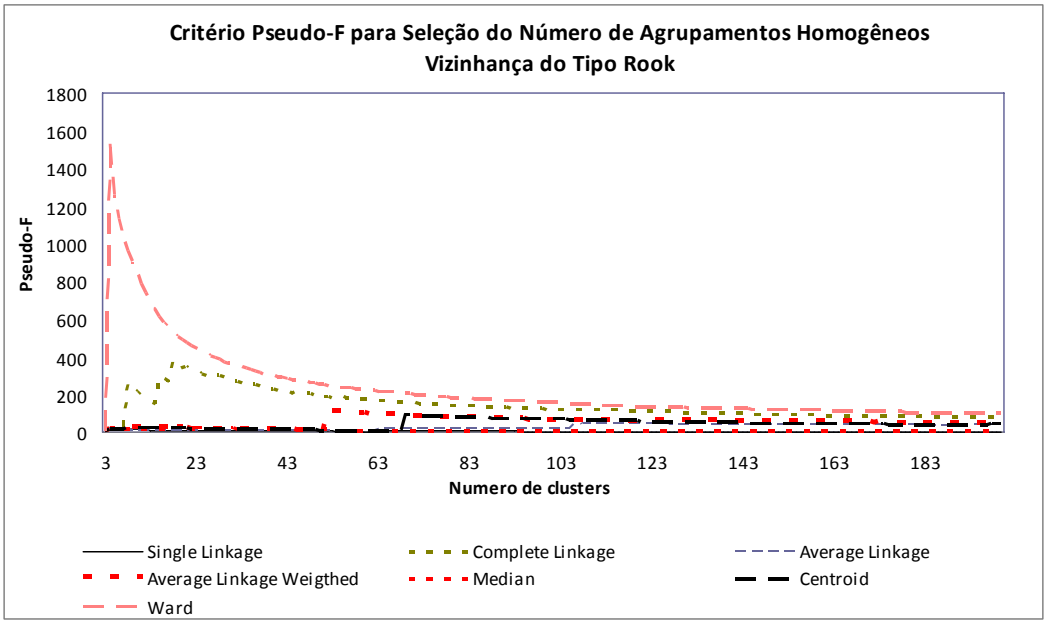
FIGURA A3.2
Critério R^2 para vizinhança do tipo *rook*



Elaboração dos autores.

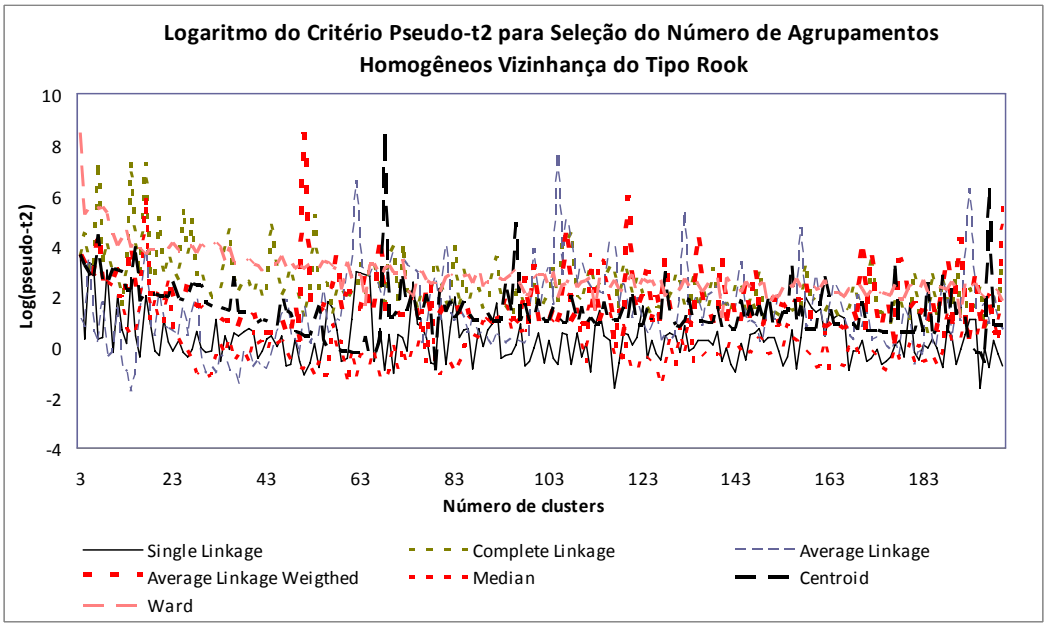
4. Para visualização em cores, acessar a seção *O trabalho do Ipea*, subseção *Publicações*, no site: <http://www.ipea.gov.br>.

FIGURA A3.3
Critério pseudo-*F* para vizinhança do tipo *rook*



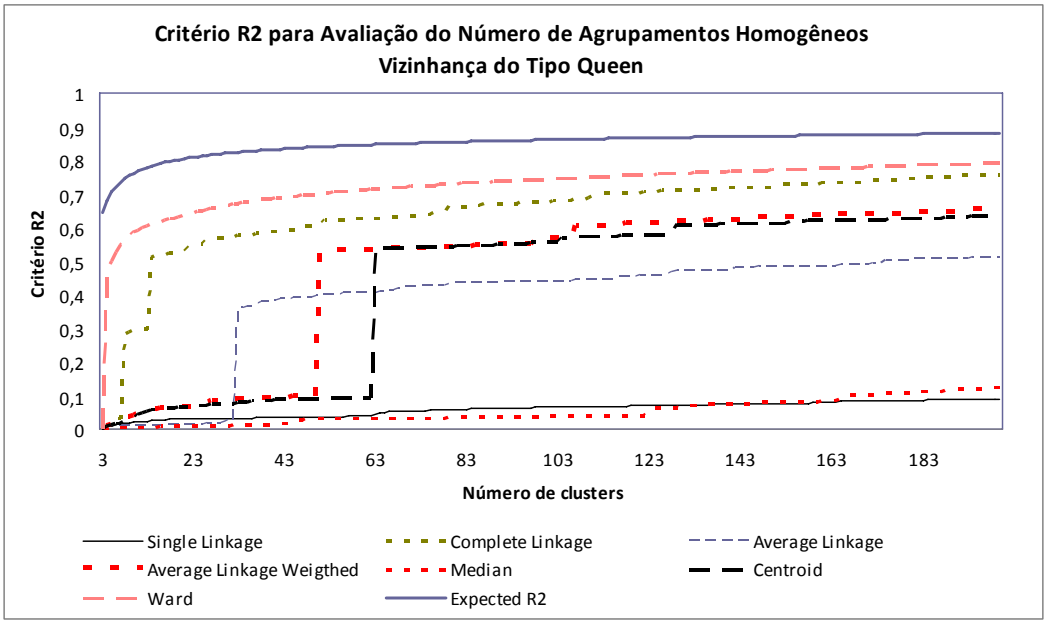
Elaboração dos autores.

FIGURA A3.4
Critério pseudo-t^2 para vizinhança do tipo *rook*



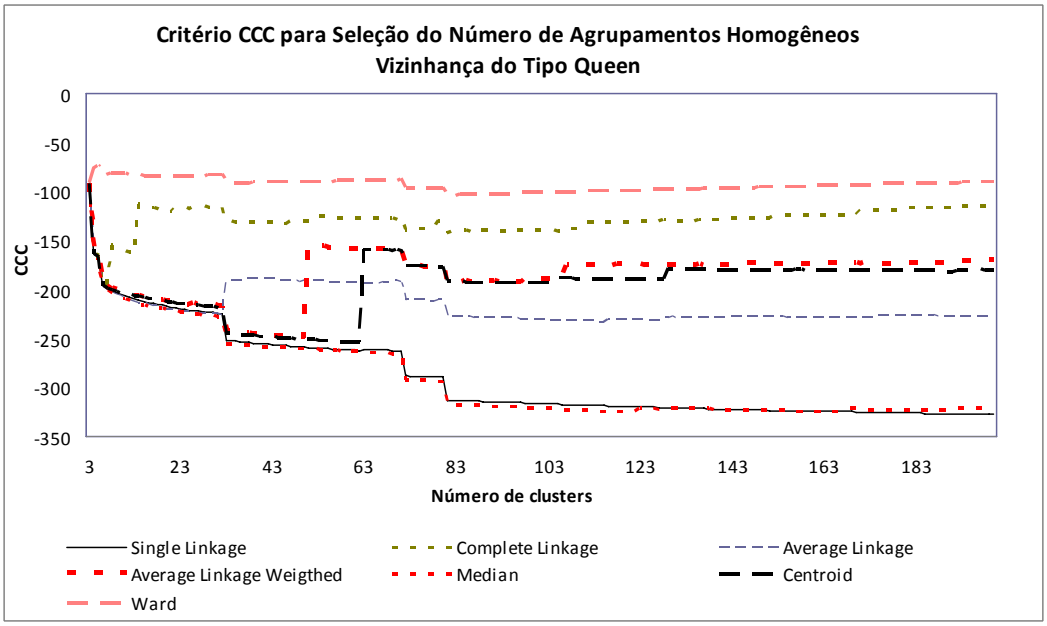
Elaboração dos autores.

FIGURA A3.5
Critério R^2 para vizinhança do tipo *queen*



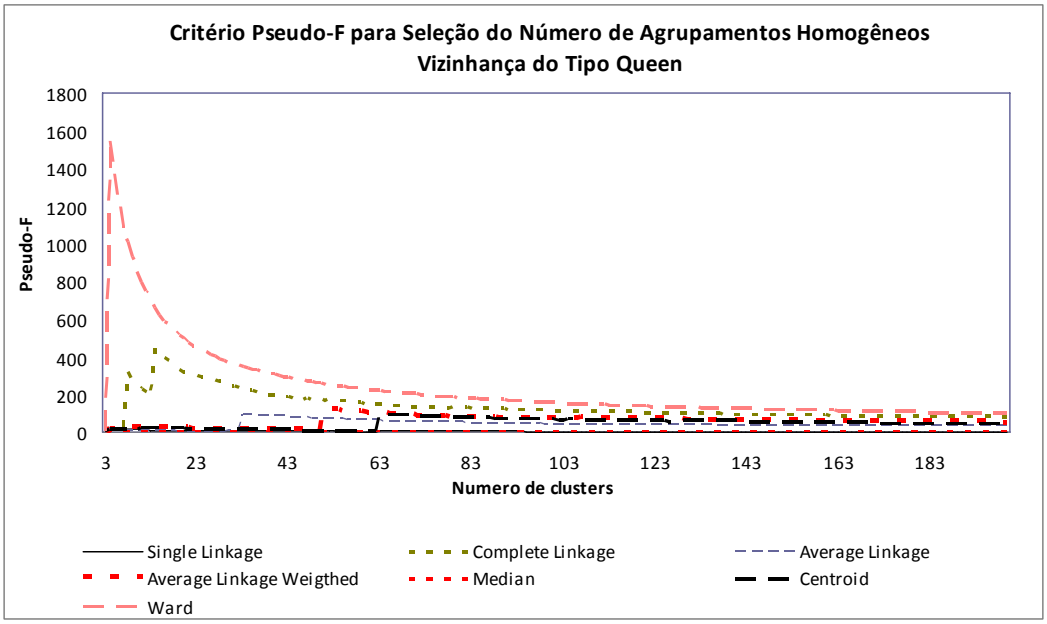
Elaboração dos autores.

FIGURA A3.6
Critério *CCC* para vizinhança do tipo *queen*



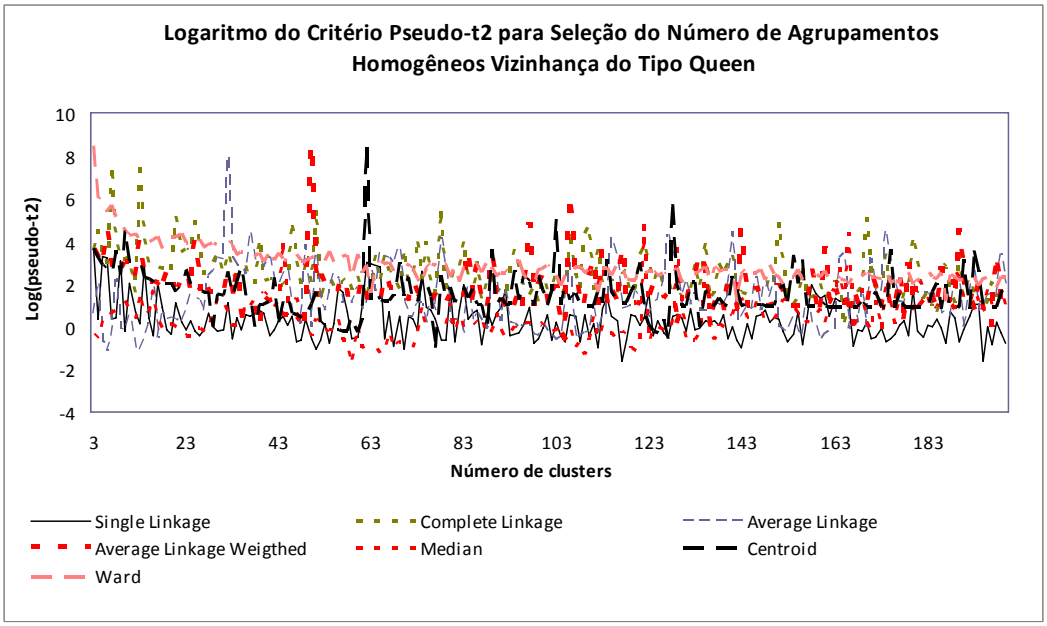
Elaboração dos autores.

FIGURA A3.7
Critério pseudo-*F* para vizinhança do tipo *queen*



Elaboração dos autores.

FIGURA A3.8
Critério pseudo-t^2 para vizinhança do tipo *queen*



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