

de Carvalho, Alexandre Ywata et al.

Working Paper

Clusterização espacial e não espacial: Um estudo aplicado à agropecuária brasileira

Texto para Discussão, No. 2279

Provided in Cooperation with:

Institute of Applied Economic Research (ipea), Brasília

Suggested Citation: de Carvalho, Alexandre Ywata et al. (2017) : Clusterização espacial e não espacial: Um estudo aplicado à agropecuária brasileira, Texto para Discussão, No. 2279, Instituto de Pesquisa Econômica Aplicada (IPEA), Brasília

This Version is available at:

<https://hdl.handle.net/10419/177495>

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

2279

TEXTO PARA DISCUSSÃO

CLUSTERIZAÇÃO ESPACIAL E NÃO ESPACIAL: UM ESTUDO APLICADO À AGROPECUÁRIA BRASILEIRA

Alexandre Xavier Ywata de Carvalho
Pedro Henrique Melo Albuquerque
Camilo Rey Laureto
Guilherme Costa Chadud Moreira
Gustavo Gomes Basso
Luiz Felipe Dantas Guimarães
Marina Garcia Pena

ipea

CLUSTERIZAÇÃO ESPACIAL E NÃO ESPACIAL: UM ESTUDO APLICADO À AGROPECUÁRIA BRASILEIRA

Alexandre Xavier Ywata de Carvalho¹

Pedro Henrique Melo Albuquerque²

Camilo Rey Laureto³

Guilherme Costa Chadud Moreira⁴

Gustavo Gomes Basso⁵

Luiz Felipe Dantas Guimarães⁶

Marina Garcia Pena⁷

1. Diretor e Técnico de planejamento e pesquisa na Diretoria de Estudos e Políticas Regionais, Urbanas e Ambientais (Dirur) do Ipea.

2. Pesquisador do Programa de Pesquisa para o Desenvolvimento Nacional (PNPD) na Dirur/Ipea.

3. Pesquisador do PNPD na Dirur/Ipea.

4. Pesquisador do PNPD na Dirur/Ipea.

5. Pesquisador do PNPD na Dirur/Ipea.

6. Pesquisador do PNPD na Dirur/Ipea.

7. Pesquisadora do PNPD na Dirur/Ipea.

Governo Federal

**Ministério do Planejamento,
Desenvolvimento e Gestão**
Ministro interino Dyogo Henrique de Oliveira

ipea Instituto de Pesquisa
Econômica Aplicada

Fundação pública vinculada ao Ministério do Planejamento, Desenvolvimento e Gestão, 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
Ernesto Lozardo

**Diretor de Desenvolvimento Institucional,
Substituto**
Carlos Roberto Paiva da Silva

**Diretor de Estudos e Políticas do Estado,
das Instituições e da Democracia**
João Alberto De Negri

Diretor de Estudos e Políticas Macroeconômicas
José Ronaldo de Castro Souza Júnior

**Diretor de Estudos e Políticas Regionais,
Urbanas e Ambientais**
Alexandre Xavier Ywata de Carvalho

**Diretora de Estudos e Políticas Setoriais
de Inovação e Infraestrutura**
Fernanda De Negri

Diretora de Estudos e Políticas Sociais
Lenita Maria Turchi

**Diretor de Estudos e Relações Econômicas
e Políticas Internacionais**
Sérgio Augusto de Abreu e Lima Florêncio Sobrinho

Assessora-chefe de Imprensa e Comunicação
Regina Alvarez

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

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.

© Instituto de Pesquisa Econômica Aplicada – **ipea** 2017

Texto para discussão / Instituto de Pesquisa Econômica Aplicada.- Brasília : Rio de Janeiro : Ipea , 1990-

ISSN 1415-4765

1. Brasil. 2. Aspectos Econômicos. 3. Aspectos Sociais.
I. Instituto de Pesquisa Econômica Aplicada.

CDD 330.908

As opiniões emitidas nesta publicação são de exclusiva e inteira responsabilidade dos autores, não exprimindo, necessariamente, o ponto de vista do Instituto de Pesquisa Econômica Aplicada ou do Ministério do Planejamento, Desenvolvimento e Gestão.

É permitida a reprodução deste texto e dos dados nele contidos, desde que citada a fonte. Reproduções para fins comerciais são proibidas.

JEL: A12; A13; C18; C38; C63; O13; R14.

SUMÁRIO

SINOPSE

1 INTRODUÇÃO	7
2 METODOLOGIA.....	8
3 BASE DE DADOS.....	18
4 RESULTADOS.....	19
5 CONSIDERAÇÕES FINAIS.....	32
REFERÊNCIAS	33
BIBLIOGRAFIA COMPLEMENTAR.....	34
APÊNDICE A	35
APÊNDICE B	41
APÊNDICE C – CATEGORIAS DAS VARIÁVEIS.....	47
APÊNDICE D – VALORES MÉDIOS DAS PRINCIPAIS VARIÁVEIS, POR <i>CLUSTER</i>	50

SINOPSE

Este trabalho apresenta uma análise de clusterização de áreas mínimas comparáveis (AMCs) para traçar um mapa de agrupamentos homogêneos a partir de uma combinação de variáveis climáticas, de características do solo e de produção agropecuária. A metodologia permite a visualização de interações entre as diversas variáveis utilizadas, identificando, por exemplo, padrões de coexistência, no nível municipal, de diferentes culturas agrícolas. A discussão apresenta os algoritmos tradicionais sem contiguidade (aglomerativo hierárquico e *k-means*) e o algoritmo aglomerativo hierárquico com imposição de contiguidade. Busca-se, dessa forma, explorar diferenças entre as tipologias construídas com diferentes abordagens, além de prover configurações alternativas de agrupamentos. As metodologias discutidas permitem ainda a incorporação de critérios tradicionais de escolha do número de *clusters*, tais como estatísticas *CCC*, *pseudo-F* e *pseudo-t²*.

Palavras-chave: clusterização espacial; variáveis agropecuárias; algoritmos hierárquicos; *k-means*.

1 INTRODUÇÃO

Áreas territoriais extensas podem implicar diversidades econômicas, culturais, sociodemográficas, comportamentais, climáticas e geográficas. Dessa forma, a formulação de políticas públicas encontra obstáculos quando implementadas nas diversas regiões do país, já que estas diversidades impõem diferentes intensidades de aplicação da política, uma vez que existem áreas mais ou menos dependentes dessas ações. O Brasil, por ser um país com grandes dimensões territoriais, pode apresentar dificuldades em implementar políticas públicas. O setor agropecuário, principalmente, sofre com essas dificuldades, já que se estende por quase todo o território nacional e é afetado com alterações nos padrões locais.

De acordo com Diniz (2013), a partir de 1970, a região de São Paulo intensificou a produção de cana de açúcar e laranja em detrimento da produção de grãos, a qual migrou para estados das regiões Sul e Centro-Oeste. Quanto ao rebanho bovino, entre 1970 e 2010, as regiões Centro-Oeste e Norte tinham 24% do rebanho nacional, porcentagem que saltou para 55% em 2010. Isso mostra o desafio que é maximizar a aplicação de políticas públicas de forma a promover equidade entre diferentes regiões geográficas.

Uma forma de solucionar esse problema é aplicar políticas por municípios, microrregiões ou mesorregiões. Entretanto, mesmo nessas divisões territoriais encontram-se heterogeneidades diversas. Um conjunto de municípios pode ter características semelhantes, não fazendo sentido analisá-los em separado. Já as micro e mesorregiões podem ser grandes o suficiente para gerar grupos territoriais com características distintas.

Uma técnica comumente utilizada para identificar grupos homogêneos de unidades observacionais é a análise de *clusters* (agrupamentos). A partir de variáveis caracterizando cada unidade, essa metodologia busca identificar subgrupos similares de observações. Com isso, é possível: *i*) a construção de conjuntos homogêneos de municípios que podem, por exemplo, receber tratamento similar em termos de políticas públicas; e *ii*) a visualização de interações entre as variáveis, caracterizando cada unidade observada. Portanto, as técnicas de *clusters* constituem uma ferramenta muito utilizada para identificação de padrões.

Este trabalho visa aplicar análises de clusterização de áreas mínimas comparáveis (AMCs) para traçar um mapa dessas áreas, contíguas e não contíguas, com características homogêneas, a partir de determinadas variáveis agropecuárias. Este estudo priorizou a utilização de AMCs, pois essa divisão territorial impõe padronização na divisão geográfica das regiões ao longo do tempo, característica não necessariamente presente nas divisões políticas municipais. Segundo Reis, Pimentel e Alvarenga (2008), entre 1872 e 2000, o número de municípios brasileiros saltou de 642 para 5.507, sendo que não necessariamente suas fronteiras foram respeitadas. Os autores definem as AMCs como um agregado de municípios passíveis de comparações intertemporais de forma significativa, não caracterizando uma divisão política ou administrativa. Dado que estamos utilizando AMCs com unidade geográfica para a análise de agrupamentos neste artigo, pode-se empregar a mesma metodologia para outros períodos de tempo, e se obterem agrupamentos comparáveis.

O trabalho está dividido em cinco seções, incluindo esta introdução. A segunda seção aborda a metodologia utilizada para formação dos grupos homogêneos de AMCs (*clusters*) – essa discussão inclui os algoritmos tradicionais sem contiguidade (aglomerativo hierárquico e *k-means*) e o algoritmo aglomerativo hierárquico com imposição de contiguidade. A terceira seção engloba a descrição dos dados utilizados na análise. A quarta seção apresenta os principais resultados oriundos da análise de *clusters* empregada. A quinta seção é reservada para as conclusões do trabalho.

2 METODOLOGIA

A clusterização de AMCs consiste no agrupamento¹ de AMCs homogêneas, sendo que esse agrupamento pode ter como pressuposto a contiguidade ou não das AMCs. Neste artigo, consideram-se algoritmos a construção de agrupamentos com e sem restrições de contiguidade entre as unidades geográficas. Busca-se, dessa forma, explorar diferenças entre as tipologias construídas com abordagens distintas, além de prover variadas configurações de agrupamentos que poderão ser utilizadas em diferentes situações de políticas públicas.

1. Agrupamentos homogêneos serão também denominados *clusters*. O algoritmo ou processo de formação de *clusters* será denominado *clusterização*.

Inicialmente, apresenta-se uma descrição sucinta dos algoritmos aglomerativos de clusterização hierárquica tradicionais e as modificações realizadas nesse método a fim de incorporar a restrição de unidades geográficas contíguas. Em seguida, discute-se um outro método para construção de *clusters* não contíguos, que não se baseia na ideia de aglomeração sequencial. Esse método é conhecido como *k-means*; uma de suas principais diferenças em relação aos métodos aglomerativos é que ele necessita, *a priori*, que o número de agrupamentos seja determinado. Por fim, são apresentados os métodos utilizados para determinar os agrupamentos de AMCs sem a restrição de contiguidade e com o pressuposto dessa restrição.

2.1 Algoritmos aglomerativos de clusterização hierárquica

Dada a grande importância do problema de clusterização de observações, pesquisadores têm se dedicado à construção de algoritmos computacionais que possam realizar automaticamente a detecção de agrupamentos homogêneos de observações – uma discussão ilustrativa é dada em Carvalho *et al.* (2011). Hastie, Tibshirani e Friedman (2001) apresentam uma descrição geral desses algoritmos, que podem ser divididos em três grandes categorias: *i*) algoritmos combinatórios (*combinatorial algorithms*); *ii*) misturas de modelos (*mixture models*); e *iii*) busca por modas (*mode seeking*). As últimas duas categorias baseiam-se em alguma forma de modelos probabilísticos para o processo gerador de dados. Já os algoritmos combinatórios podem ser vistos basicamente como regras heurísticas de busca dos melhores agrupamentos de observações. De maneira geral, não existe um algoritmo que seja superior aos demais em todas as situações.

Este trabalho segue a dinâmica dos algoritmos combinatórios, os quais têm uma estrutura de formação aglomerativa de *clusters* do tipo hierárquica, conforme trabalho de Khattree e Naik (2000). De maneira geral, essa estrutura segue os seguintes passos.

- 1) Seja uma base de N *clusters* iniciais. Em geral, esses agrupamentos correspondem simplesmente às unidades a serem agrupadas em grupos homogêneos (por exemplo, municípios). Portanto, em geral, cada um desses N *clusters* contém apenas uma unidade inicialmente. A cada unidade i está associado um vetor de m características $x_i = [x_{i,1} \ x_{i,2} \ \dots \ x_{i,m}]$. Essas características podem ser socioeconômicas, por exemplo.
- 2) Calcula-se a distância entre todos os pares formados por elementos entre esses N *clusters* iniciais. Distância, nesse caso, pode ser qualquer métrica de dissimilaridade

(ou similaridade, dependendo do algoritmo) entre o conjunto de atributos $x_i = [x_{i,1} \ x_{i,2} \ \dots \ x_{i,m}]$. Para uma discussão sobre as diversas métricas de dissimilaridade e similaridade, ver Khattree e Naik (2000) e Berry e Linoff (1997). Entre as diversas medidas de dissimilaridade possíveis, pode-se citar a medida de Ward, que permite encontrar *clusters* de tamanhos não muito diferentes² e pode ser escrita como:

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

Onde $D_{K,L}$ é a medida de dissimilaridade (ou distância) entre o *cluster* L e o *cluster* K , \bar{x}_L e \bar{x}_K são os vetores correspondentes às médias dos vetores de características de todas as unidades (por exemplo, municípios) dentro dos *cluster* L e K , respectivamente, e N_L e N_K são as quantidades de unidades dentro dos *clusters* L e K . A função $\|\cdot\|$ corresponde à norma euclidiana. Outras medidas de dissimilaridade podem ser consultadas em Carvalho *et al.* (2009; 2011) e Charrad *et al.* (2014).

- 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 fim do processo, ter-se-á em mãos uma árvore descrevendo a sequência de agrupamentos em cada passo do algoritmo, sendo essa representação por vezes denominada dendograma. Para um número inicial de N unidades observacionais na base de dados, ao todo ocorrem $N - 1$ junções. Diversos *softwares* estatísticos possuem recursos gráficos que permitem ao usuário apresentar a árvore construída.

2. Conforme exposto em Carvalho *et al.* (2009; 2011), a medida de Ward permite gerar *clusters* com menor variabilidade total, tornando-os assim mais homogêneos. Utilizando-se outros critérios de dissimilaridade, o número de elementos dos *clusters* pode variar de forma mais significativa.

Algoritmos aglomerativos hierárquicos em geral são muito demandantes computacionalmente. Na primeira iteração do processo, o número de pares de observações possíveis é igual a $N \times (N - 1) / 2$. 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 torna-se impossível. Nessas situações, diversas alternativas existem, como o sorteio de uma subamostra das N observações para posterior comparação.

Para situações envolvendo unidades geográficas, como municípios ou setores censitários, no entanto, 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 descrito a seguir, em cada iteração do algoritmo, o número de pares de observações comparadas não mais será $N \times (N - 1) / 2$, para o caso de comparações feitas apenas entre unidades geográficas vizinhas. Isso reduz enormemente o tempo de processamento.

O passo final é então selecionar o número de *clusters* ou de grupos homogêneos. Diversas medidas estatísticas para seleção do número de agrupamentos foram propostas, sem haver necessariamente um consenso sobre qual medida utilizar. Algumas dessas estatísticas são a *CCC*, a *pseudo-F* e a *pseudo-t²* (ver Khattree e Naik, 2000). De maneira geral, essas medidas estão associadas a um indicador de dissimilaridade agregada entre todos os *clusters* construídos.

Por meio de um gráfico das medidas citadas *versus* o número de *clusters* selecionado, é possível identificar aumentos expressivos (picos) no grau de dissimilaridade para algum número específico de *clusters*. Esses picos no grau de dissimilaridade agregada sugerem então os pontos de parada no algoritmo de agregação sequencial apresentados nos passos 1 a 5 descritos anteriormente, indicando, portanto, quantos *clusters* utilizar. Por sua vez, é interessante ter uma interpretação plausível para todos os *clusters* formados, pois permite combinar algoritmos computacionais robustos com o escopo do estudo.

2.2 Algoritmos de clusterização hierárquica espacial

No contexto deste trabalho, as unidades a serem agrupadas são AMCs, que formarão *clusters* de AMCs homogêneas, para as quais políticas de desenvolvimento regional

específicas possam ser propostas. Nesse caso, espera-se que os *clusters* formados agreguem AMCs homogêneas e espacialmente vizinhas.

Já a aplicação direta de qualquer um dos algoritmos comumente encontrados na literatura e disponíveis em pacotes estatísticos, muito provavelmente fornecerá grupos homogêneos formados por AMCs que não apresentem contiguidade entre elas. Dessa forma, foram propostas algumas modificações no algoritmo de clusterização, a fim de incorporar explicitamente a restrição de contiguidade entre as AMCs que compõem um mesmo *cluster*, conforme a seguir.

- 1) Seja C uma base inicial de N unidades geográficas, que já podem corresponder a agrupamentos iniciais de subunidades (neste texto, essas subunidades correspondem a AMCs). Inicialmente, cada uma dessas N observações consiste em um *cluster* isoladamente e tem um conjunto de atributos $[x_{i,1} \ x_{i,2} \ \dots \ x_{i,m}]$. Para cada uma dessas N unidades, encontra-se a lista de vizinhos, de acordo com algum critério espacial. Nesse caso, foram definidas como vizinhas as AMCs que contêm pelo menos um lado (ou dois pontos) em comum, num sistema de georreferenciamento. Esse tipo de contiguidade é conhecido como contiguidade do tipo *rook* (Anselin, 1988; Anselin e Florax, 2000).³
- 2) Calcula-se a distância entre todos os pares formados por elementos estritamente vizinhos na lista de N unidades, segundo medida de Ward descrita no passo 2 do algoritmo hierárquico tradicional. O número de pares testados nesse caso 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, reduzindo assim consideravelmente o tempo de processamento.
- 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 $[x_{i,1} \ x_{i,2} \ \dots \ x_{i,m}]$ mas também as listas de vizinhos. Portanto, será composta uma nova lista de AMCs vizinhas a partir da união da lista de vizinhos da AMC I com a lista de vizinhos da AMC J .

3. Alternativamente, poderíamos ter escolhido a vizinhança do tipo *queen*. Nesse tipo de vizinhança, duas unidades geográficas são consideradas vizinhas caso elas tenham pelo menos um ponto em comum. Em Carvalho *et al.* (2009; 2011), os autores identificaram que a utilização de vizinhança do tipo *queen* pode incorrer em *clusters* que, apesar de contíguos, apresentam formas muito irregulares. Por esse motivo, optou-se pela utilização de vizinhança do tipo *rook*.

- 5) Para os $N - 1$ novos *clusters*, depois da junção descrita nos itens 3 e 4, calculam-se as distâncias entre todos os pares de *clusters* vizinhos. Nesse caso, dois *clusters* A e B de AMCs são considerados vizinhos quando houver pelo menos uma AMC em A que é vizinho de uma AMC 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 *clusters* A e B corresponde unicamente à dissimilaridade entre os atributos $[x_{i,1} \ x_{i,2} \ \dots \ x_{i,m}]$. Em geral, essas variáveis correspondem a características socioeconômicas e não contêm necessariamente informações sobre localização geográfica. A similaridade geográfica já está explicitamente modelada quando são agrupados somente *clusters* vizinhos.
- 6) Repetem-se os passos 2 a 5 até se obter um único *cluster*, que deverá conter todas as N unidades geográficas originais.

Da mesma forma que no caso da clusterização hierárquica tradicional, ao fim do processo, tem-se uma árvore caracterizando os agrupamentos decorridos em cada passo do algoritmo. Novamente, pode-se recorrer a alguns dos indicadores tradicionais (por exemplo, *CCC*, *pseudo-F* e *pseudo-t²*) para a escolha do número de agrupamentos mais apropriado. Para definição dos *clusters*, utilizam-se critérios objetivos, subjetivos ou a combinação deles.

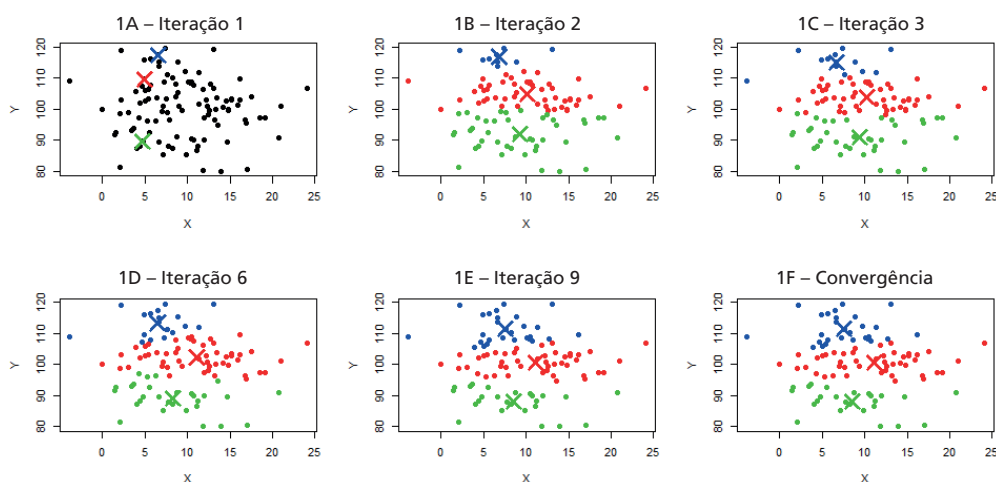
Uma explicação a respeito dos diversos critérios de seleção do número de *clusters*, comumente utilizados em *softwares* estatísticos, é apresentada em Charrad *et al.* (2014). A escolha do número de agrupamentos homogêneos via critérios subjetivos foi utilizada, por exemplo, em Chein, Lemos e Assunção (2005), na qual foram selecionados cem *clusters* para todo o território brasileiro. Para esse estudo, utilizaram-se critérios objetivos, com análises a partir de inspeções visuais, para detecção do número de *clusters* de AMCs homogêneas contíguas.

2.3 Algoritmo de agrupamentos *k-means*

Os algoritmos aglomerativos hierárquicos correspondem a um dos tipos de metodologias para se obter conjuntos homogêneos de observações em uma base de dados. Há vários outros algoritmos na literatura. Um dos métodos utilizados é o algoritmo *k-means*. Essa metodologia não possui um caráter sequencial, da mesma forma que os algoritmos hierárquicos, ela possui um caráter iterativo para o qual temos que especificar o número de agrupamentos *a priori*.

O gráfico 1 apresenta uma ilustração dos passos no algoritmo *k-means*. Inicialmente, parte-se de uma base de dados com um conjunto de variáveis, as quais pretende-se encontrar grupos similares (no gráfico 1, temos apenas duas variáveis, X e Y). Especifica-se então, *a priori*, um número de agrupamentos para identificação e para cada um desses agrupamentos, seleciona-se uma semente para inicialização do algoritmo. Para mais detalhes sobre a escolha das sementes *vide* Hastie, Tibshirani e Friedman (2001).

GRÁFICO 1
Ilustração do algoritmo *k-means* para encontrar os grupos homogêneos



Elaboração dos autores.

Obs.: Figura reproduzida em baixa resolução em virtude das condições técnicas dos originais (nota do Editorial).

Inicialmente são escolhidas três sementes, uma vez que o número de *clusters* especificado, *a priori*, é igual a três. Essas três sementes são os pontos verde, vermelho e azul da iteração 1 no gráfico 1A.

O próximo passo é determinar, para cada observação da amostra, qual a semente mais próxima. Nesse caso, a noção de proximidade pode ser qualquer métrica genérica de dissimilaridade. A mais utilizada é a distância euclidiana. Outras distâncias, entretanto, podem ser incorporadas de acordo com a necessidade da aplicação e da natureza das variáveis. Em Carvalho *et al.* (2009) são apresentadas algumas distâncias para o caso de variáveis quantitativas, enquanto em Carvalho *et al.* (2011), para variáveis binárias.

Na iteração 2, gráfico 1B, note que todos os pontos da amostra foram coloridos em verde, vermelho ou azul, dependendo de qual semente é a mais próxima. Calcula-se então o centroide de todos os pontos em cada cor. Esses centroides passam, portanto, a ser as novas sementes.

A partir das novas sementes, o algoritmo encontra novamente, para cada ponto na amostra, qual a semente mais próxima. Nesse caso, a atribuição de cores é refeita, de acordo com as outras iterações, e novos centroides são encontrados, redefinindo-se as sementes. Esse processo iterativo procede até haver convergência. O algoritmo converge quando a atribuição de cores, devido às alterações de centroides, não mais acontece.

Dado que é necessário especificar o número de agrupamentos de forma antecipada, o algoritmo *k-means* não permite o levantamento de indicadores para sugestão do número de agrupamentos a serem utilizados. Por sua vez, conforme veremos mais adiante, para um mesmo número de agrupamentos, o algoritmo *k-means* resulta em medidas globais de homogeneidade da clusterização melhores do que o algoritmo aglomerativo hierárquico.

Por esses motivos, o procedimento geral foi utilizar o algoritmo aglomerativo hierárquico para sugerir o número de *clusters*, e definido esse número de agrupamentos, executou-se o algoritmo *k-means*.

Na próxima subseção, faremos uma discussão de métricas indicativas da homogeneidade dos *clusters* gerados.

2.4 Medidas de homogeneidade para os *clusters* formados

Uma das medidas a serem realizadas entre os agrupamentos contíguos e os não contíguos é comparar o número de agrupamentos no caso contíguo para se atingir o mesmo nível de homogeneidade dos agrupamentos não contíguos.

Em geral, quando impomos a restrição de contiguidade, necessitamos de um maior número de agrupamentos para obtermos o mesmo grau de homogeneidade. Para estudar a diferença entre o número de agrupamentos nos casos contíguos e não contíguos, para se atingir o mesmo nível de homogeneidade, consideraremos o critério

R^2 . Dentro de cada agrupamento k , considere a variabilidade quadrática dada por WSS_k (*within sum of squares*), calculada como:

$$WSS_k = \sum_{i=1}^{n_k} \sum_{j=1}^p (x_{i,j} - \bar{x}_{k,j})^2 \quad (2)$$

Onde n_k é o número de AMCs no *cluster* k ; p é o número de variáveis para gerar os *clusters*; $x_{i,j}$ é o valor da variável j , na AMC i ; e $\bar{x}_{k,j}$ é a média da variável j para as AMCs no *cluster* k . A variabilidade total é calculada pelo somatório das variabilidades dentro de cada *cluster*. Portanto, a variabilidade total (WSS) pode ser obtida pela expressão:

$$WSS = \sum_{k=1}^K WSS_k \quad (3)$$

No caso de apenas um *cluster*, a variabilidade total quadrada é denominada somatório total dos quadrados, TSS (*total sum of squares*), e tem expressão:

$$TSS = \sum_{i=1}^n (x_{i,j} - \bar{x}_{.,j})^2 \quad (4)$$

Onde n corresponde ao número total de AMCs e $\bar{x}_{.,j}$ corresponde à média global da variável j . A medida R^2 é então dada pela expressão:

$$R^2 = 1 - \frac{WSS}{TSS} \quad (5)$$

Pode-se mostrar que o coeficiente R^2 varia de 0 a 1, sendo $R^2 = 0$ quando o número de *clusters* é igual a 1 e $R^2 = 1$ quando o número de *clusters* é igual a n . Quanto mais homogêneos forem os *clusters*, menores serão os valores WSS_k . Se os valores das variáveis nas AMCs em cada *cluster* fossem exatamente os mesmos, teríamos $WSS_1 = WSS_2 = \dots = WSS_K = 0$, e nesse caso $WSS = 0$, resultando em $R^2 = 1$. Portanto, quanto mais próximo o R^2 for de 1, mais homogêneos são os *clusters* gerados.

2.5 Agrupamentos sem a restrição de contiguidade

Os métodos que não impõem contiguidade podem ser mais interessantes quando o objetivo é destacar a grandeza de cada variável. Isso se deve ao fato de a contiguidade imposta resultar em grupos menos homogêneos, ou, de outra forma, grupos em que nem todas as unidades apresentam características tão semelhantes, mas foram unidos por serem vizinhos.

A formação de agrupamentos homogêneos de municípios, com componentes não necessariamente contíguos, pode não ser um problema em muitas das aplicações. De fato, pode acontecer de um analista estar interessado justamente em identificar se existem regiões que são semelhantes em termos de atributos socioeconômicos, mas a contiguidade não ser a condição preponderante.

Para a construção de agrupamentos sem imposição da contiguidade, utilizaram-se algoritmos tradicionais, do tipo *k-means*, e algoritmos aglomerativos hierárquicos, conforme detalhadamente discutido nas subseções anteriores.

2.6 Agrupamentos com a restrição de contiguidade

A intuição do argumento para clusterização de AMCs é que, por hipótese, as variáveis utilizadas serão suficientes para descrever as características das AMCs estudadas. Pode acontecer, no entanto, 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 para o agrupamento de AMCs.

Dessa forma, espera-se que as variáveis ausentes na base de dados apresentem uma forte correlação espacial global, no sentido de que AMCs vizinhas tenham características semelhantes (Anselin, 1988; Anselin e Florax, 2000; Pace e Barry, 1997). Nesse caso, a utilização de algoritmos de clusterização em que a contiguidade é imposta pode reduzir a perda de informação devido à ausência de algumas variáveis na base.

Para construir os agrupamentos de AMCs homogêneas de forma contígua, este trabalho utiliza a metodologia apresentada em Carvalho *et al.* (2009; 2011), definida a partir de modificações nos algoritmos tradicionais de clusterização hierárquica, conforme discutido nas subseções 2.1 a 2.4.

Algoritmos alternativos para clusterização espacial – ou seja, com *clusters* compostos por unidades contíguas – estão descritos 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 expostas em Assunção, Lage e Reis (2002) e Chein, Lemos e Assunção (2005).

Neste trabalho, entretanto, optou-se por modificar os algoritmos de clusterização hierárquica tradicionais, descritos em Khattree e Naik (2000) e Berry e Linoff (1997). O objetivo é constituir uma abordagem mais intuitiva, na qual os passos dos algoritmos ficam nítidos tanto para os usuários quanto para os leitores. Em adição, a nova metodologia permite a incorporação imediata de diferentes medidas de dissimilaridade entre grupos homogêneos e critérios tradicionais de escolha do número de *clusters*.

3 BASE DE DADOS

Neste estudo foram utilizadas as bases de dados de 2012 da Produção Agrícola Municipal (PAM) e da Produção da Pecuária Municipal (PPM), ambas elaboradas pelo Instituto Brasileiro de Geografia e Estatística (IBGE) e disponibilizadas em seu portal para consulta. A PAM compreende as culturas de maior relevância na lavoura brasileira, tanto em produção quanto no comércio. A PPM agrupa dados do setor pecuário para os municípios brasileiros, por exemplo, número de bovinos, suínos e caprinos.

Além disso, também foi utilizado um conjunto de bases de dados provenientes do IBGE com o objetivo de retratar as características do solo e do clima para o território brasileiro, conforme a seguir.

- 1) Características físicas: considera os aspectos do solo que podem influenciar no desenvolvimento das plantas, como armazenamento de água e aeração do solo.
- 2) Tipos de clima: descreve o nível de umidade conforme o número de meses com baixa umidade.
- 3) Fertilidade: analisa os atributos químicos do solo que podem ajudar ou restringir o desenvolvimento das plantas, por exemplo, a concentração de nutrientes benéficos e de alumínio.

- 4) Limitações do solo: considera as limitações naturais de cada tipo de solo, como a disponibilidade de nutrientes e a característica de relevo.
- 5) Seca: analisa a duração e a distribuição da seca ao longo dos meses.
- 6) Tipos de solo: agregado de fatores que têm por objetivo avaliar o potencial do solo para uso agrícola.
- 7) Temperatura: classifica as áreas de acordo com as temperaturas médias.
- 8) Topologia: analisa e classifica o terreno de acordo com as ondulações e declives da superfície.
- 9) Uso do solo: classifica o terreno de acordo com a sua utilização.

Uma descrição detalhada das categorias utilizadas pode ser consultada no apêndice C.

A utilização desses bancos gerou um conjunto de 118 variáveis de análise, número elevado para construir um modelo e com grande probabilidade de alta correlação entre elas. Por esse motivo, foi efetuada uma análise de componentes principais para redução de dimensionalidade do banco de dados.

A análise de componentes principais permite, por exemplo, acomodar situações nas quais diversas variáveis possuem alta correlação entre elas. Selecionamos um número de componentes que contabilizasse 99% da variabilidade das 118 variáveis originais. Ao final, selecionamos 83 componentes – que foram então utilizados como variáveis para as análises de agrupamentos.⁴ Essa análise é melhor detalhada em Hastie, Tibshirani e Friedman (2001).

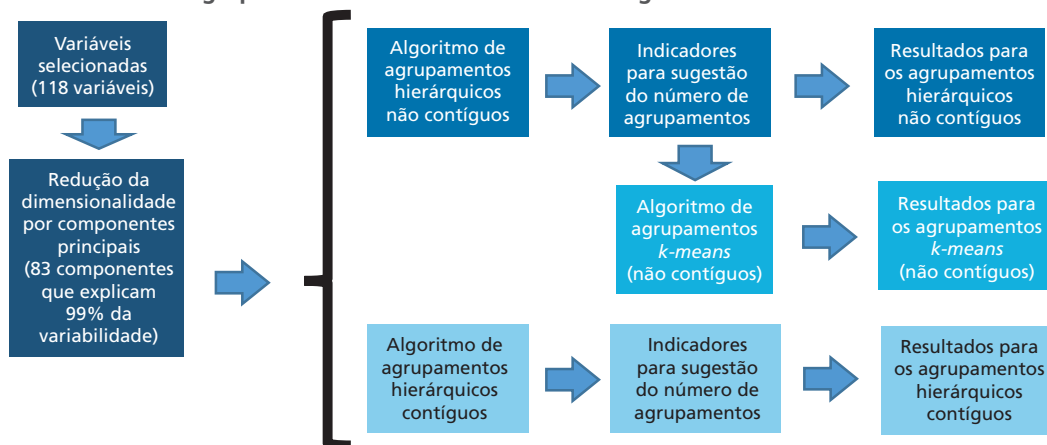
4 RESULTADOS

A figura 1 apresenta os passos para as análises de agrupamentos, considerando-se um conjunto de 83 componentes principais físicos, climáticos e agropecuários, conforme descrito na seção 3. Com base nestes 83 componentes, foram efetuadas análises de agrupamentos considerando-se três metodologias: clusterização hierárquica não espacial, *k-means* e clusterização hierárquica espacial.

4. Todos os cálculos foram efetuados utilizando-se o *software* livre R, e os mapas foram elaborados nos *softwares* QuantumGis ou ArcGis.

FIGURA 1

Análise de agrupamentos com diferentes metodologias



Elaboração dos autores.

4.1 Identificação do número de *clusters*

A primeira metodologia abordada foi a análise de agrupamentos aglomerativos hierárquicos não contíguos. Para isso, utilizaram-se os critérios *CCC*, *pseudo-F* e *pseudo-t²* a fim de determinar o número de *clusters*. Nos dados em análise, a medida *pseudo-t²* foi a que melhor elucidou o número de *clusters* a ser utilizado. Ao longo da sequência aglomerativa de *clusters*, pelo método hierárquico, a medida *pseudo-t²* indica junções de dois *clusters* que são relativamente diferentes entre si, ou seja, a medida *pseudo-t²* aponta situações de junção forçada de dois grupos de AMCs. Sugere-se então utilizar o número de agrupamentos imediatamente anterior à junção dos grupos não muito semelhantes.

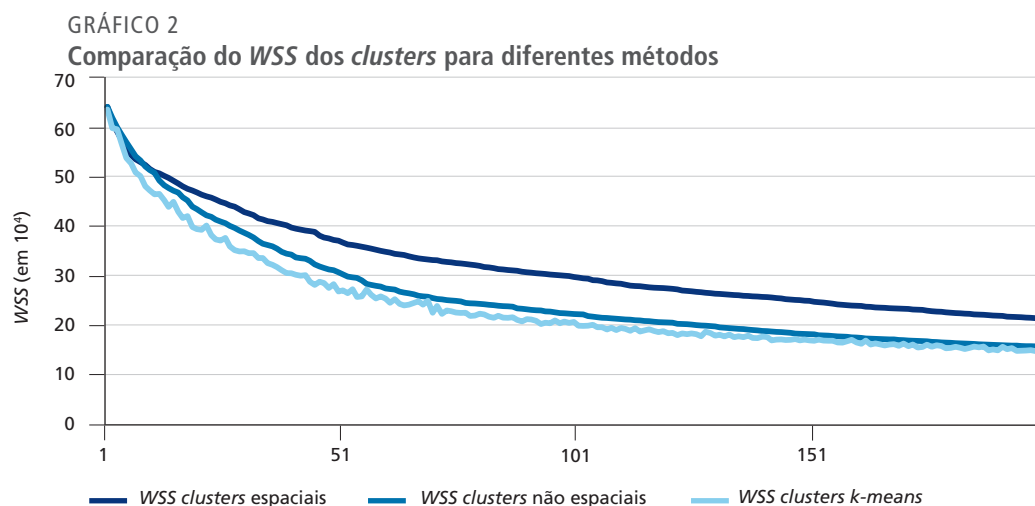
Em geral, na sequência de aglomerações, a indicação de junções forçadas aparece em diversos pontos. Isso significa que podemos selecionar números diferentes de *clusters*, mesmo usando apenas um único critério (no caso, o *pseudo-t²*). Neste artigo, procuraram-se junções forçadas em quatro pontos da sequência aglomerativa, o que resultou na indicação de quatro números diferentes de *clusters* a serem utilizados. Os números sugeridos pela estatística *pseudo-t²* foram 6, 8, 16 e 21 *clusters*.

A segunda metodologia baseou-se no método *k-means*, o qual considerou a mesma metodologia de agrupamentos sugeridos na primeira análise. Sendo assim, em

ambos os métodos não espaciais, os números de agrupamentos considerados foram os mesmos, com diferença apenas na composição de cada *cluster*.

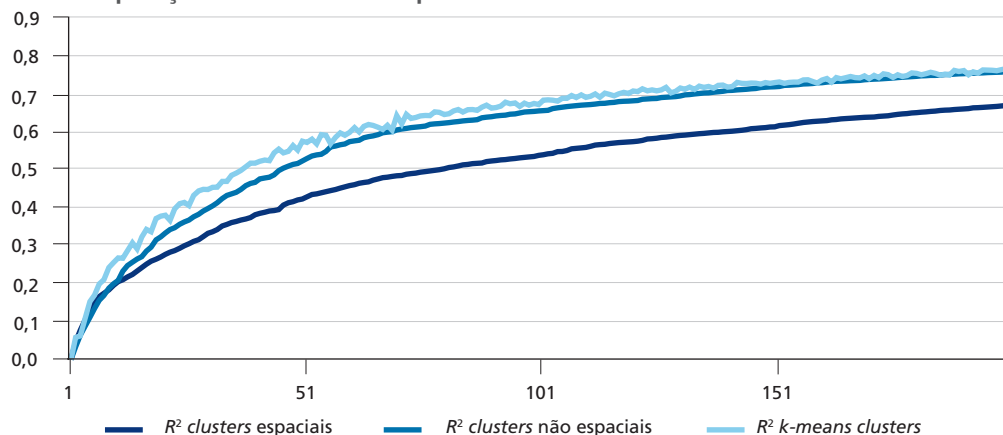
Por fim, na terceira metodologia, que utiliza o método de clusterização hierárquica considerando contiguidade espacial, o resultado apresentou os seguintes números de *clusters*: 40, 47, 52 e 65. O motivo de se utilizar uma maior quantidade de agrupamentos para o caso contíguo é explicado pelo fato de os *clusters* espaciais implicarem restrições de contiguidade, isto é, o algoritmo penaliza as AMC's não contíguas, podendo formar, em geral, uma maior quantidade de *clusters* com menos AMC's em cada. Logo, para se obter um nível equivalente de homogeneidade agregada dos agrupamentos não espaciais, um maior número de *clusters* é requerido pelos agrupamentos espaciais.

Os gráficos 2 e 3 exibem, respectivamente, a variabilidade total (WSS) e o R^2 em relação ao número de *clusters*, para os três métodos discutidos anteriormente. Estes gráficos têm por objetivo ilustrar um comparativo entre as estatísticas de cada método e auxiliar na interpretação dos resultados.



Elaboração dos autores.

GRÁFICO 3
 Comparação do R^2 dos *clusters* para diferentes métodos



Elaboração dos autores.

A partir de dez agrupamentos, os *clusters* esaciais possuem menor R^2 para um mesmo número de *clusters*. Antes disso, o R^2 para o método hierárquico não espacial era menor que para o método hierárquico espacial. Para o método *k-means*, no entanto, o R^2 só é menor para quatro *clusters* e maior para todos os outros casos. Isto é esperado em casos gerais, uma vez que a espacialidade limita os agrupamentos entre as AMCs e, portanto, tende a diminuir o R^2 . Observamos, por exemplo, que o R^2 para o método *k-means*, com 21 *clusters*, é igual a aproximadamente 38%. Este mesmo valor aproximado pode ser observado, no método espacial hierárquico, para quarenta *clusters*.

Pode-se notar também que, entre os métodos utilizados, o *k-means* apresenta maior R^2 . Um dos motivos para isso é o fato de o método hierárquico ser sequencial e, portanto, depender, a cada passo, de agregação dos passos dados anteriormente. Além disso, o algoritmo hierárquico é do tipo *greedy*, ou seja, junções em um determinado passo não levam em conta quais serão as junções futuras. Ainda, destaca-se o fato de a curva R^2 não ser estritamente crescente para o caso *k-means*, que ocorre devido ao fato de este método não ser sequencial, ao contrário dos dois métodos hierárquicos.

Entre os quatro valores para o número de *clusters* sugeridos, decidiu-se utilizar o valor igual a 21 *clusters* para os casos não esaciais. Além de ser o valor apontado pelas estatísticas no critério de parada, esse número fornece também análises mais interessantes pelas características de homogeneidade em cada *cluster* gerado, de acordo com as variáveis selecionadas.

Para o método espacial, optou-se por utilizar o número igual a quarenta *clusters*. Novamente, este número foi sugerido pelos critérios de parada e, além disso, este caso possui R^2 bastante próximo àquele encontrado com 21 *clusters* para o método *k-means*, permitindo assim certo grau de comparação entre os métodos.

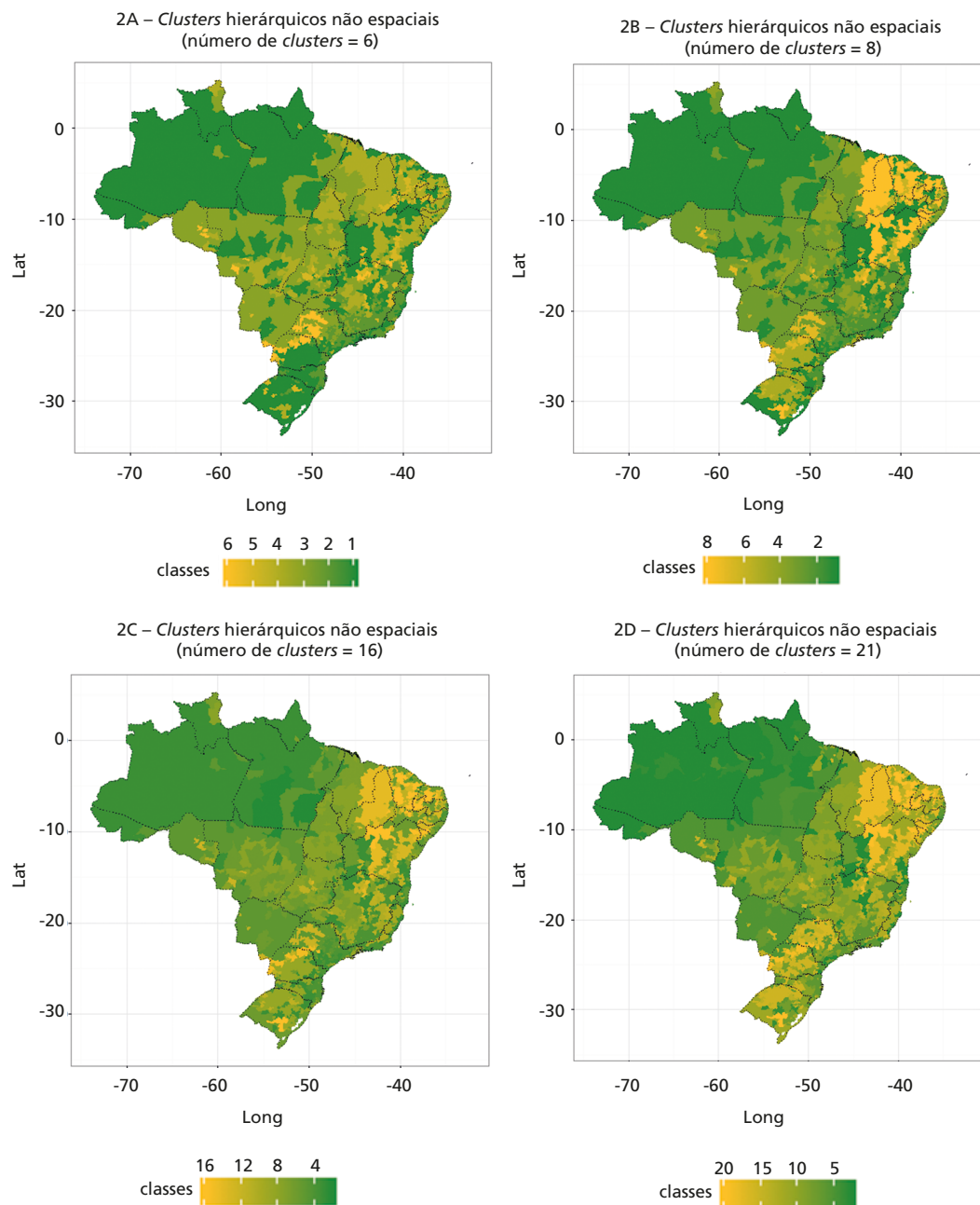
4.2 Resultados para *clusters* por métodos não espaciais

As figuras a seguir apresentam os números de *cluster* sugeridos e seus respectivos mapas para cada um dos métodos. Como mencionado anteriormente, tanto para o método aglomerativo hierárquico não espacial quanto para o método *k-means* foi considerado o mesmo número de agrupamentos. Dessa forma, nos mapas, é possível visualizar a composição dos *clusters* formados em cada um dos métodos, assim como sua disposição geográfica no território nacional. As cores ilustram os *clusters* pela porcentagem de floresta presente em cada mapa, com a cor mais amarelada significando uma menor e a mais esverdeada uma maior porcentagem de floresta.

Nas análises dos resultados, o foco foi mais concentrado no método *k-means*, devido ao fato de os agrupamentos deste método terem apresentado um maior R^2 . Estas análises são feitas a partir dos *box-plots* (diagramas de caixa) presentes no apêndice A, de forma a obter os quartis de cada elemento, em cada *cluster* e por variável. Os quartis centrais, formados pelos valores entre 25% e 75%, mostram os valores não discrepantes da variável dentro de cada *cluster*. Portanto, decidiu-se utilizar a mediana, como medida de centralidade, por ela ser robusta a observações extremas dentro de cada *cluster*. É possível, ainda, consultar, no apêndice D, os valores médios das principais culturas e criações analisadas por *cluster*, assim como o número de AMCs por *cluster*.

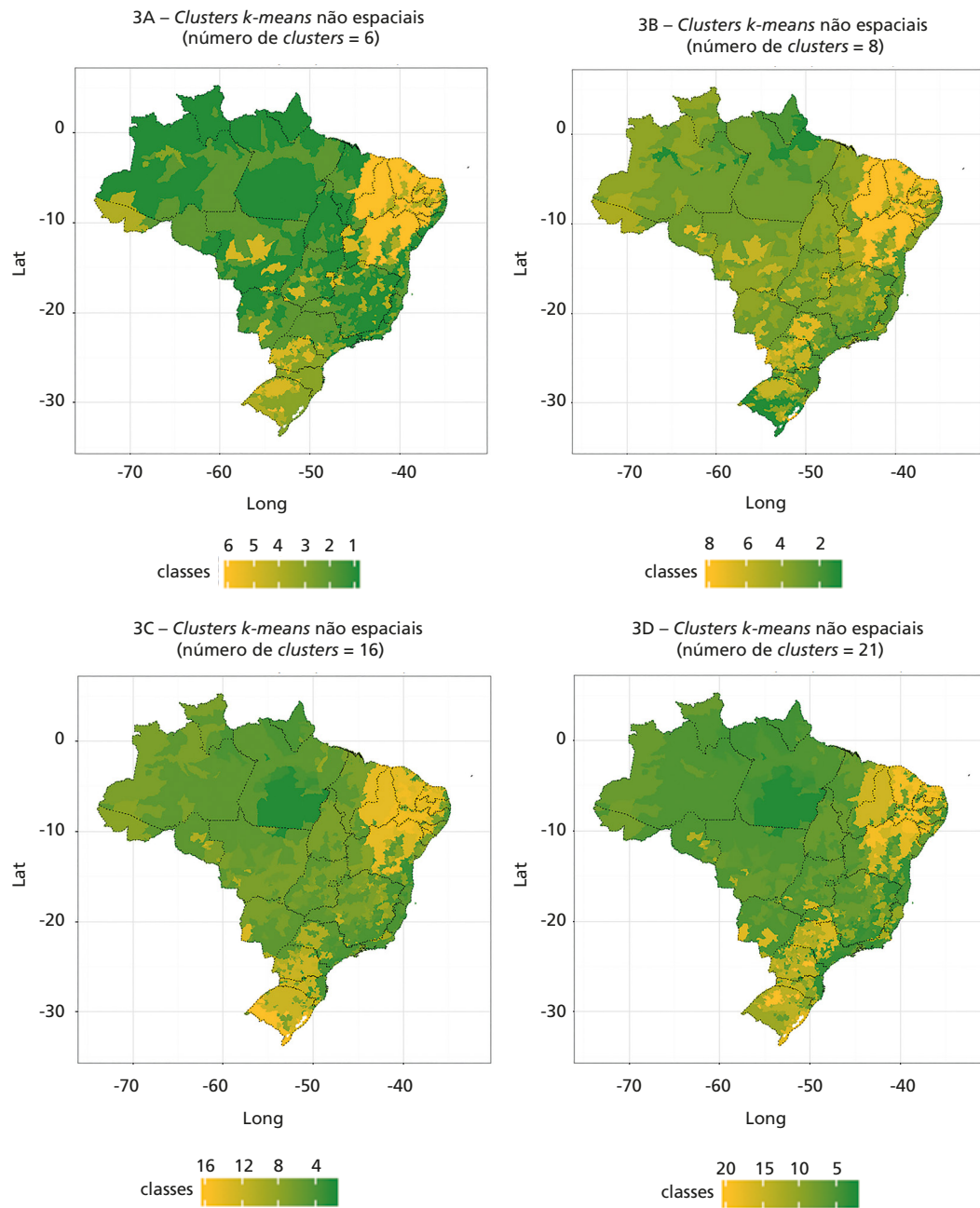
Entre os *clusters* para o método *k-means*, o *cluster* 5 se destaca, em termos de mediana, em algumas das principais criações e culturas do setor agropecuário, como as criações de bovinos, suínos, ovinos e equinos, enquanto na agricultura são relevantes as culturas de sorgo, algodão, soja, milho, feijão e arroz.

FIGURA 2
Clusters não contíguos com o método aglomerativo hierárquico



Elaboração dos autores.

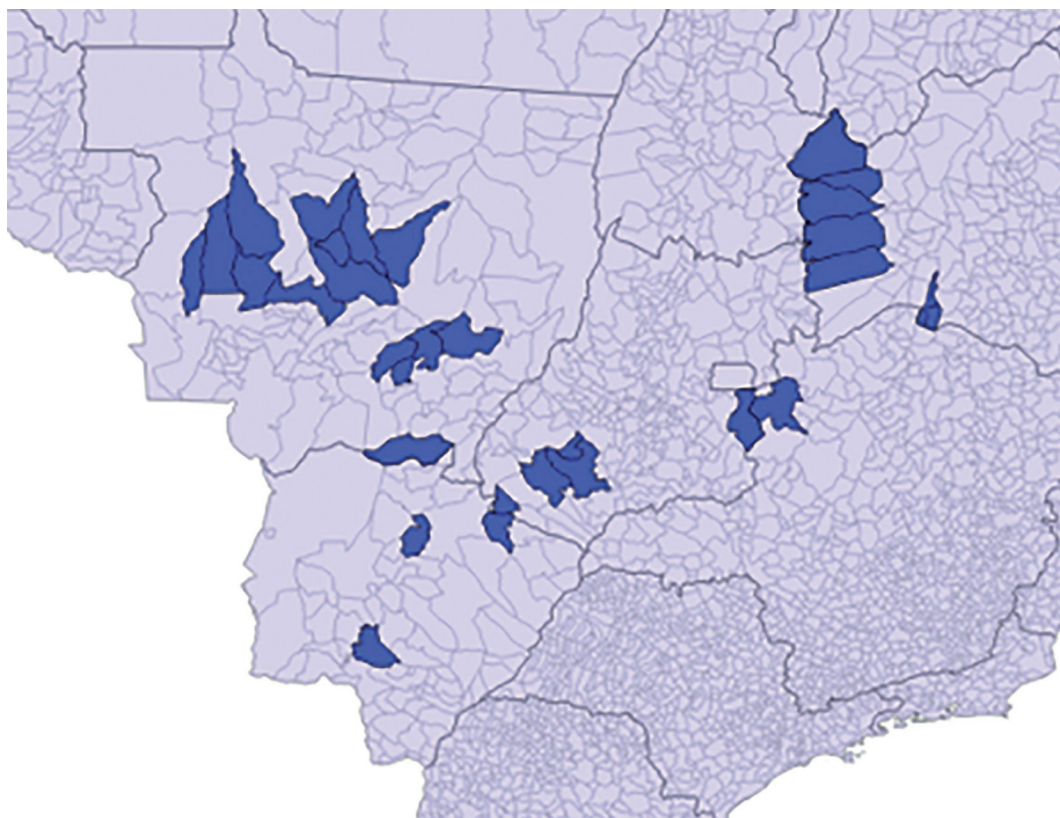
FIGURA 3
Clusters não contíguos com o método k-means



Elaboração dos autores.

A figura 4 exibe a posição do *cluster* 5 no território brasileiro. Nota-se que esse *cluster* é formado por AMCs presentes nos estados do Mato Grosso, Mato Grosso do Sul, Goiás, Minas Gerais e Bahia. Apresenta, ainda, dominância de fertilidade baixa, temperatura quente e subquente, clima semiúmido e uso do solo voltado para área agrícola e pastagem plantada.

FIGURA 4
Cluster número 5 – método *k-means*



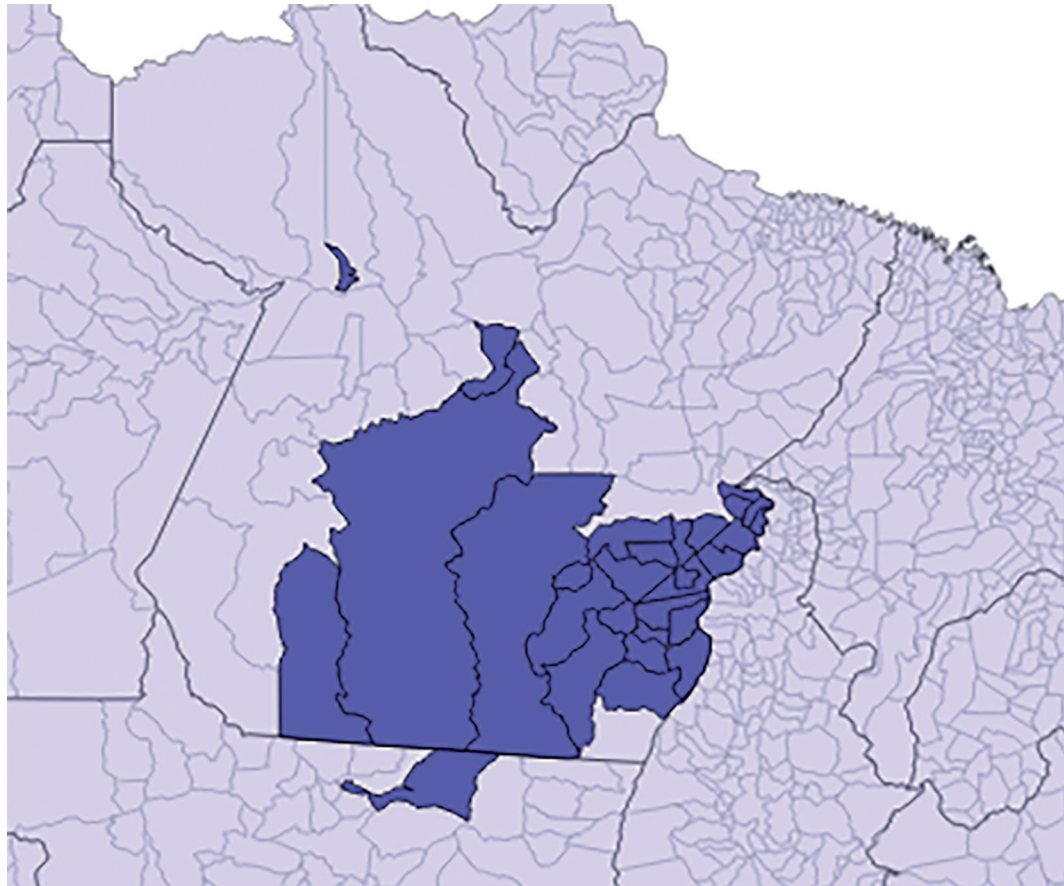
Elaboração dos autores.

Obs.: Figura reproduzida em baixa resolução em virtude das condições técnicas dos originais (nota do Editorial).

Outro *cluster* que se destaca é o *cluster* 1 (figura 5). Primeiramente, apesar de não se tratar de um método espacial, as unidades que o compõem são bastante próximas, apresentando contiguidade entre todas as AMCs, menos uma. Na produção agrícola, as culturas de maior destaque foram arroz e mandioca, enquanto, na pecuária, as principais criações foram as de bovinos e equinos.

Como pode ser observado na figura 5, o agrupamento 1 localiza-se, primordialmente, no estado do Pará. Suas características climáticas são temperatura quente e subquente e clima úmido. Com relação ao solo, sua fertilidade é média e o seu uso é destinado principalmente à área de vegetação florestal e à pastagem plantada.

FIGURA 5
Cluster número 1 – método k-means



Elaboração dos autores.

Obs.: Figura reproduzida em baixa resolução em virtude das condições técnicas dos originais (nota do Editorial).

Para as demais análises, novamente, o estudo focou-se nos *clusters* que possuem maior mediana em cada cultura ou criação entre as que ainda não foram analisadas. Para o setor de pecuária, o *cluster* 15 apresentou maior mediana na criação de suínos e galináceos. Em relação aos ovinos, o *cluster* 19 exibiu maior mediana, seguido do *cluster* 18, que, por sua vez, possui o maior valor máximo. Similarmente, para a criação

de caprinos, o *cluster* 16 apresentou maior mediana seguido pelo *cluster* 20, que, neste caso, apresentou o valor máximo. No que se refere ao setor de agricultura, o *cluster* 21 se destaca na produção de trigo, com maior mediana e valor máximo de produção. O *cluster* 17 apresentou maior destaque na produção de cana-de-açúcar e de laranja, enquanto os *clusters* 8, 10 e 19 se destacaram, respectivamente, na produção de mandioca, café e castanha-de-caju.

Com os resultados obtidos nesta subseção, pode-se observar, por meio dos *box-plots*, o comportamento dos *clusters* em relação à produção no setor agropecuário e, assim, diversas outras análises podem ser feitas, como a relação entre criações e culturas e as características físicas e climáticas de cada *cluster*.

4.3 Resultados para *clusters* por métodos espaciais

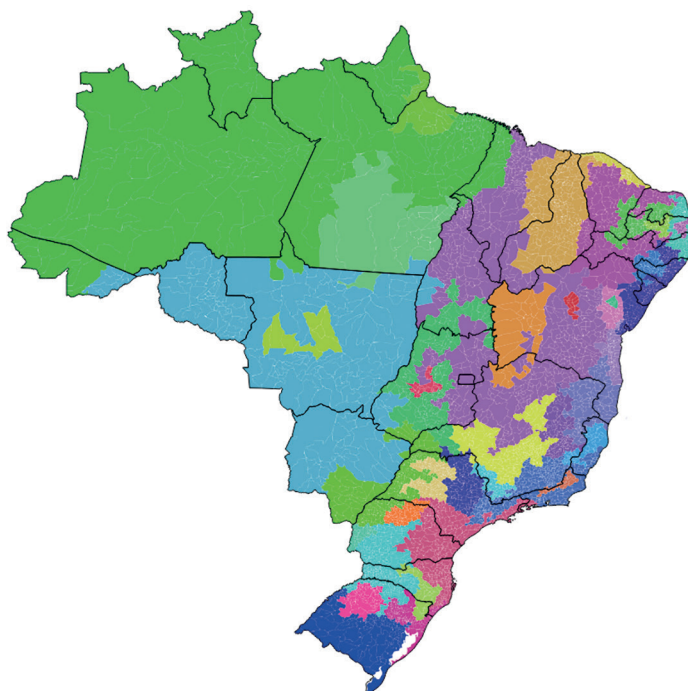
Para o caso de clusterização espacial, considera-se, no modelo, a contiguidade entre as AMCs. A presença de contiguidade torna endógenas ao modelo questões geográficas, além de levar em consideração as características descritas nas variáveis. Dessa forma, como o algoritmo penaliza unidades não contíguas e diferenças muito grandes nas características das AMCs, pode ocorrer a formação de *clusters* pequenos. Com o algoritmo hierárquico espacial, metodologia encontrada em Carvalho *et al.* (2009; 2011), os *clusters* contíguos são definidos para utilização nesta análise. Entre os números de *clusters* sugeridos pelos indicadores, optou-se por utilizar quarenta agrupamentos, já que o R^2 nesse modelo é bastante próximo do R^2 para o método *k-means* com 21 *clusters*, discutido na subseção anterior.

A figura 6 refere-se aos quarenta *clusters* encontrados para o método espacial. Como o número de *clusters* é relativamente grande, a análise por mapas com cores gradativas, como aqueles da subseção anterior, fica com difícil visualização. Dessa forma, utilizamos cores aleatórias no mapa para facilitar a visibilidade de contiguidade.

No apêndice B, estão presentes os *box-plots* para o caso de clusterização espacial, que apresentam uma análise gráfica das principais culturas e criações levadas em consideração no estudo. Dessa forma, pode-se observar por meio da mediana (medida de centralidade) os principais valores não discrepantes das variáveis dentro de cada *cluster* analisado. Ademais, assim como para os métodos não espaciais, no apêndice D

encontram-se os valores médios das principais culturas e criações analisadas por *cluster* e o número de AMCs por *cluster*.

FIGURA 6
Clusters – método espacial



Elaboração dos autores.

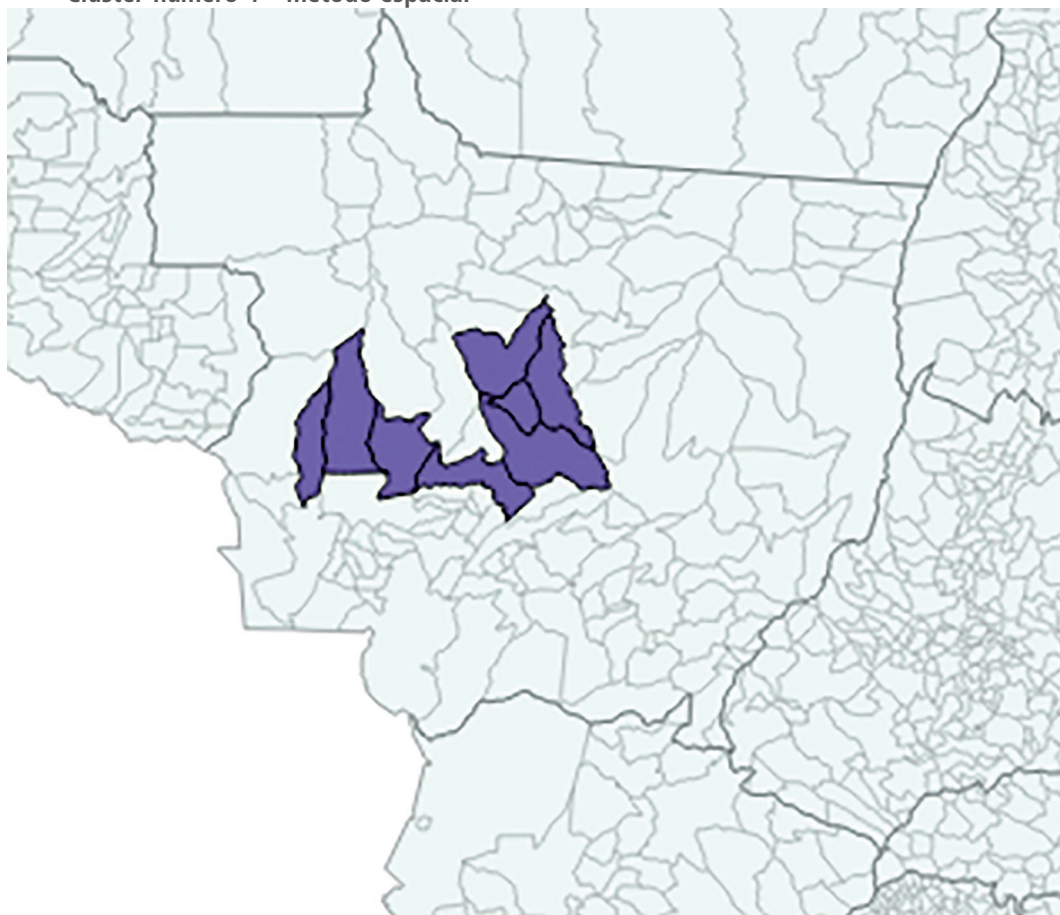
Entre os principais agrupamentos para o método espacial, o *cluster 4* é o que mais se destaca no setor agropecuário, possuindo maior mediana em diversas culturas e criações. No caso específico da agricultura, este *cluster* possui números expressivos nas produções de: *i*) soja, sendo o agrupamento de maior relevância; *ii*) milho, novamente o de maior mediana; *iii*) feijão, maior mediana, seguido pelos *clusters 24* e *21*; *iv*) arroz, com mediana maior, porém próxima à do *cluster 31*, que por sua vez possui maior valor máximo entre todos; *v*) algodão, sendo o de maior relevância; e *vi*) sorgo, mais uma vez a única mediana a se destacar em relação aos outros *clusters*.

Outra característica relevante deste *cluster* é a sua alta participação na pecuária, ressaltando-se a criação de suínos e galináceos. Entretanto, diferentemente do caso analisado para agricultura, na pecuária, as criações se mostraram com medianas mais

próximas entre os agrupamentos. Em relação às características de clima e solo, também expostas nos *box-plots*, observa-se que as principais finalidades do solo, em termos de mediana, são as áreas de vegetação florestal e as áreas agrícolas com predominância de baixa fertilidade nessa última. O clima, por sua vez, mostrou-se predominantemente semiúmido e com temperatura quente e subquente.

A figura 7 exhibe a disposição do *cluster* no Brasil, sendo ele localizado totalmente no Mato Grosso. Por ser um agrupamento que demonstrou alta relevância nas culturas e criações, destacam-se as AMCs que compõem esse *cluster*: Campo Novo do Parecis; Campos de Júlio; Diamantino; Lucas do Rio Verde; Nova Mutum; Sapezal; Sorriso; e Tapurah.

FIGURA 7
Cluster número 4 – método espacial

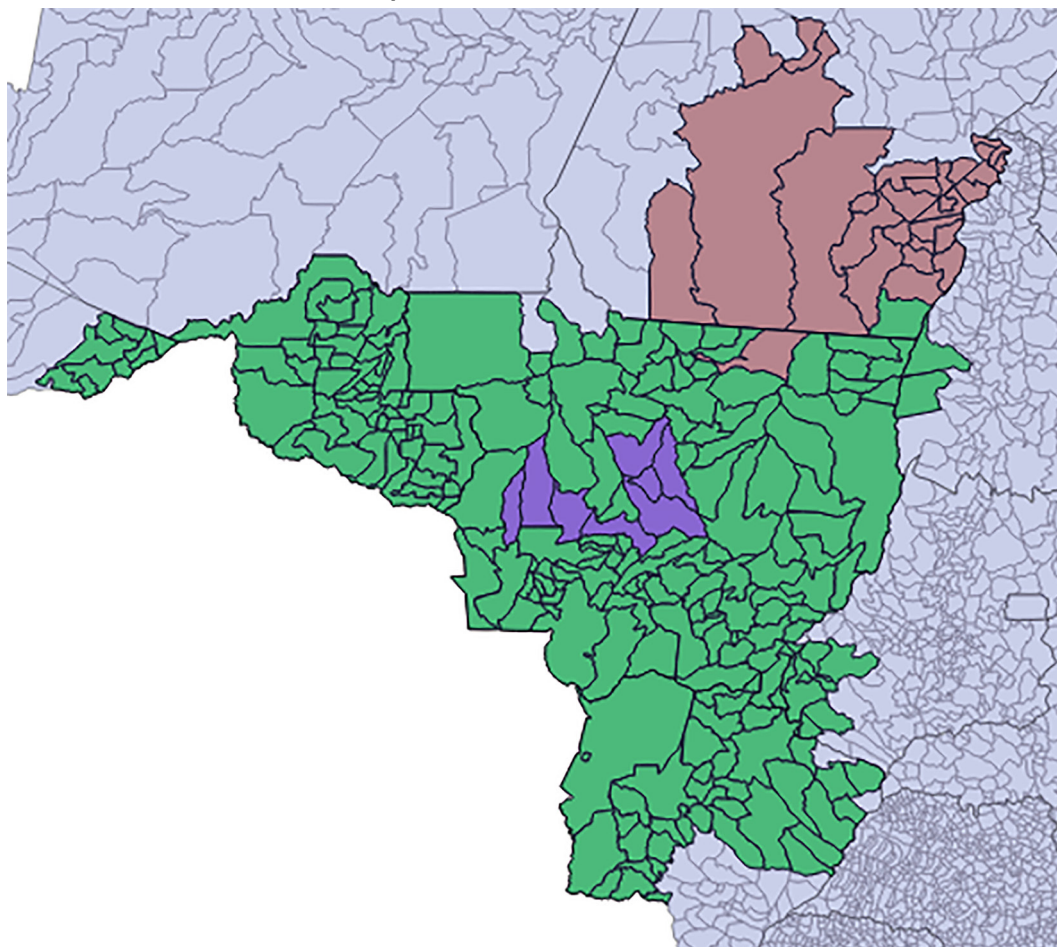


Elaboração dos autores.

Obs.: Figura reproduzida em baixa resolução em virtude das condições técnicas dos originais (nota do Editorial).

É interessante notar que o *cluster* 4 é o destaque em produção de sorgo e de milho, ambos utilizados na composição de ração usada na alimentação bovina. Por sua vez, os *clusters* 6 e 3 (em verde e marrom, respectivamente, na figura 8) são, justamente, os destaques, em termos de mediana, na criação de bovinos. Subentende-se que os *clusters* que produzem matéria-prima para a produção de um bem, como a ração para o gado, podem estimular o desenvolvimento de *clusters* próximos que se beneficiem da produção dessa matéria-prima.

FIGURA 8
Clusters 3, 4 e 6 – método espacial



Elaboração dos autores.

Obs.: Figura reproduzida em baixa resolução em virtude das condições técnicas dos originais (nota do Editorial).

Ademais, cabe ressaltar os *clusters* que apresentaram maior mediana nas culturas e nas criações que não foram mencionadas. Na agricultura, o *cluster* 20 apresentou maior

mediana na produção de cana-de-açúcar, seguido pelos *clusters* 37 e 38. Para a cultura de café, o *cluster* de maior destaque foi o *cluster* 10 e, em seguida, o *cluster* 13. Em relação à mandioca, esta foi a cultura que apresentou medianas mais próximas entre os agrupamentos analisados, possuindo como maior relevância os *clusters* 1, 5, 21, 35 e 3. O *cluster* 23 apresentou maior destaque na cultura de trigo. No setor de pecuária, para a criação de bovinos, os *clusters* de maior destaque foram o 3 e o 6. Os agrupamentos 28, 25 e 36 apresentaram as maiores medianas para caprinos. Em relação à criação dos ovinos, os *clusters* 36, 27 e 28 foram os de maior relevância. Já na criação de equinos, foram os *clusters* 3, 6 e 35, tendo os demais medianas relativamente próximas.

Os resultados anteriores, utilizando o método hierárquico com contiguidade, possibilitam aos analistas informações úteis acerca do padrão espacial investigado. Especificamente neste trabalho, alguns padrões interessantes do setor agropecuário brasileiro são identificados, permitindo, assim, a proposição de políticas públicas específicas para cada conglomerado segundo suas características intrínsecas. Uma maior ênfase foi dada ao *cluster* 4 por ter se destacado em uma quantidade considerável de culturas e criações. A relação obtida entre os *clusters* 4, 6 e 3 demonstra que futuros estudos podem ser desenvolvidos utilizando a clusterização espacial, assim como outras análises podem ser realizadas observando a relação entre as variáveis do setor agropecuário e as características de solo e clima em cada *cluster*.

5 CONSIDERAÇÕES FINAIS

O objetivo deste artigo foi efetuar a identificação de agrupamentos de AMCs para a análise de variáveis de produção agropecuária, físicas e climáticas, possibilitando serem traçadas diretrizes para o desenvolvimento de políticas públicas. O estudo realizado por meio de agrupamentos traz diversas vantagens para a análise dos dados, facilitando sua observação e minimizando problemas de dimensionamento, que podem ocorrer tanto quando uma área de estudo é suficientemente grande e com características muito genéricas, de forma que não reflita a realidade observada, quanto pequena demais, impossibilitando a identificação de padrões mais abrangentes.

No estudo, foram adotados os métodos *k-means*, aglomerativo hierárquico não espacial e o aglomerativo hierárquico espacial. Por meio de uma análise do *pseudo-t²*, do *CCC* e do *pseudo-F*, o resultado obtido foi uma seleção de quatro configurações de *clusters*

com tamanhos coincidentes para os métodos que não impõem contiguidade, e outras quatro configurações para o método espacial. Como o método *k-means* apresentou consistentemente um R^2 maior que o método aglomerativo não espacial, optou-se por uma análise conjunta do método *k-means* com o método hierárquico espacial.

Para o método *k-means*, além de uma visão geral de sua aplicação, ressaltou-se o grupo de AMCs que inclui Unaí (Minas Gerais), Nova Mutum e Sorriso (Mato Grosso) e Rio Verde (Goiás), e mais 27 AMCs, que compõem o *cluster 5*, pelo fato de este *cluster* apresentar maior mediana em algumas das culturas consideradas, em especial soja, milho, feijão, arroz, algodão e sorgo. Para a clusterização espacial, o grupo de AMCs incluindo os municípios de Mato Grosso como Campo Novo do Parecis, Campos de Júlio, Diamantino, Lucas do Rio Verde, Nova Mutum, Sapezal, Sorriso e Tapurah recebeu mais foco. Ao observar os mapas contendo os *clusters*, nos deparamos com o fato de as AMCs do agrupamento incluindo Nova Mutum, Sapezal, Sorriso e Tapurah para o método espacial estarem contidas no agrupamento incluindo Unaí, Nova Mutum, Sorriso e Rio Verde para o método *k-means*. Por um lado, este ponto reafirma o conceito de clusterização; por outro, evidencia as consequências decorrentes da imposição da necessidade de contiguidade.

Pôde-se, portanto, de maneira mais completa, exibir as distribuições das culturas agrícolas, e suas relações com características físicas e climáticas, ajudando em suas análises de alocação. Com isso, este artigo coloca em foco questões agropecuárias para as AMCs brasileiras, de forma a servir como guia para políticas públicas de desenvolvimento da agropecuária, podendo levar a políticas públicas direcionadas e especializadas. Por sua vez, utilizando-se outros conjuntos de variáveis, as metodologias discutidas aqui poderão ser aplicadas a outros temas de interesse para análise.

REFERÊNCIAS

- ANSELIN, L. **Spatial econometrics: methods and models**. Dordrecht: Kluwer Academic, 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**, v. 63, n. 220, p. 7-24, 2002.

BERRY, M.; LINOFF, G. **Data mining techniques**: for marketing, sales, and customer support. New Jersey: John Wiley and Sons, 1997.

CARVALHO, A. X. Y. *et al.* Spatial hierarchical clustering. **Revista Brasileira de Biometria**, São Paulo, v. 27, n. 3, p. 412-443, 2009.

_____. Clusterização hierárquica espacial com atributos binários. **Revista Brasileira de Biometria**, São Paulo, v. 19, n.1, p.147-197, 2011.

CHARRAD, M. *et al.* Nbclust: an R package for determining the relevant number of clusters in a data set. **Journal of Statistical Software**, v. 61, n. 6, 2014.

CHEIN, F.; LEMOS, M. B.; ASSUNÇÃO, J. J. Desenvolvimento desigual: evidências para o Brasil. *In*: ENCONTRO NACIONAL DE ECONOMIA, 33., 2005, Natal, Rio Grande do Norte. **Anais...** Natal: Anpec, 2005.

DINIZ, C. C. **Dinâmica regional e ordenamento do território brasileiro**: desafios e oportunidades. Belo Horizonte: Cedeplar/UFMG, 2013. (Texto para Discussão, n. 471).

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

KHATTREE, R.; NAIK, D. N. **Multivariate data reduction and discrimination with SAS software**. Cary: SAS Institut, 2000.

MARAVALLE, M.; SIMEONE, B. A spanning three heuristic for regional clustering. **Communications in Statistics – Theory and Methods**, v. 24, p. 629-663, 1995.

MARAVALLE, M.; SIMEONE, B.; NALDINI, R. Clustering on trees. **Computational Statistics and Data Analysis**, New York, v. 24, p. 217-234, 1997.

PACE, K.; BARRY, R. Sparse spatial autoregressions. **Statistics and Probability Letters**, v. 33, p. 291-297, 1997.

REIS, E.; PIMENTEL, M.; ALVARENGA, A. I. **Áreas mínimas comparáveis para os períodos intercensitários de 1872 a 2000**. Rio de Janeiro: Ipea/Dimac, 2008. Mimeografado.

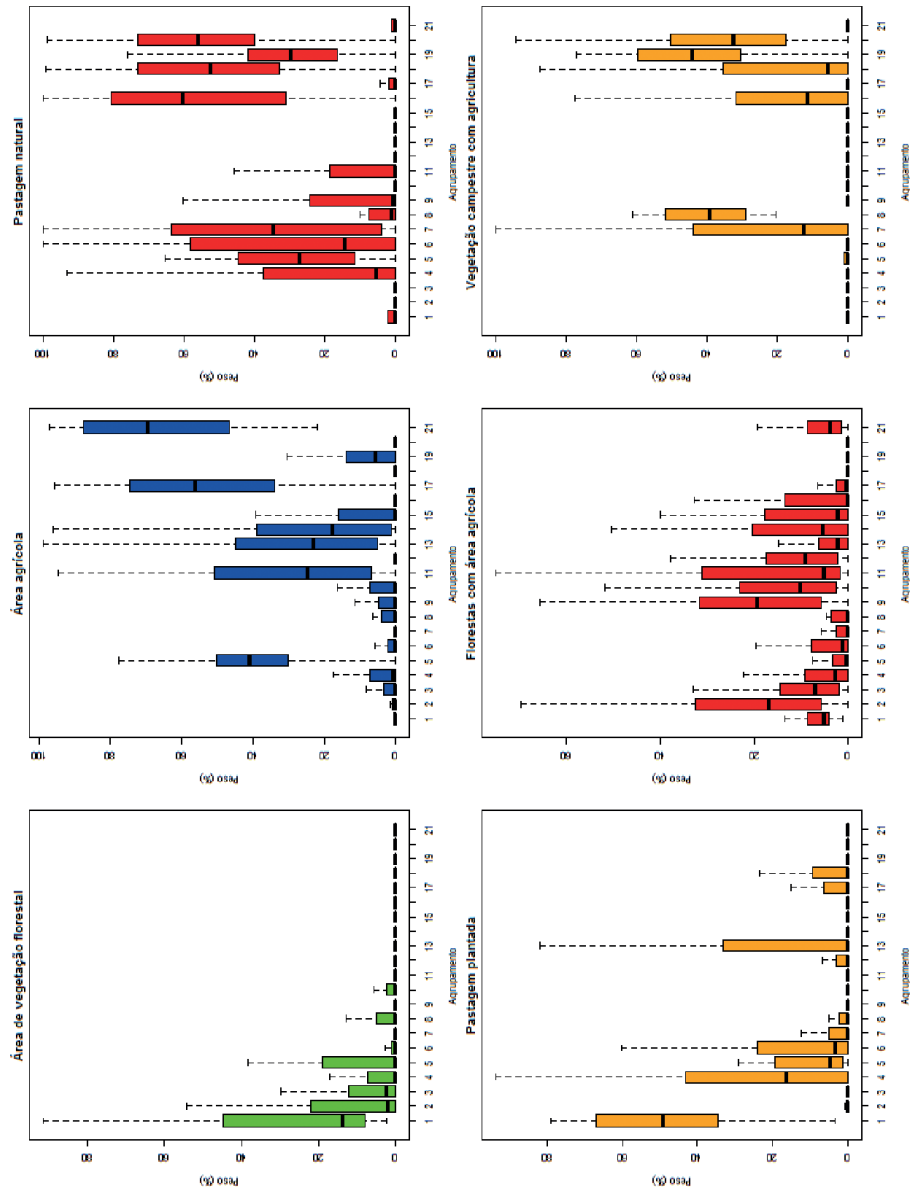
BIBLIOGRAFIA COMPLEMENTAR

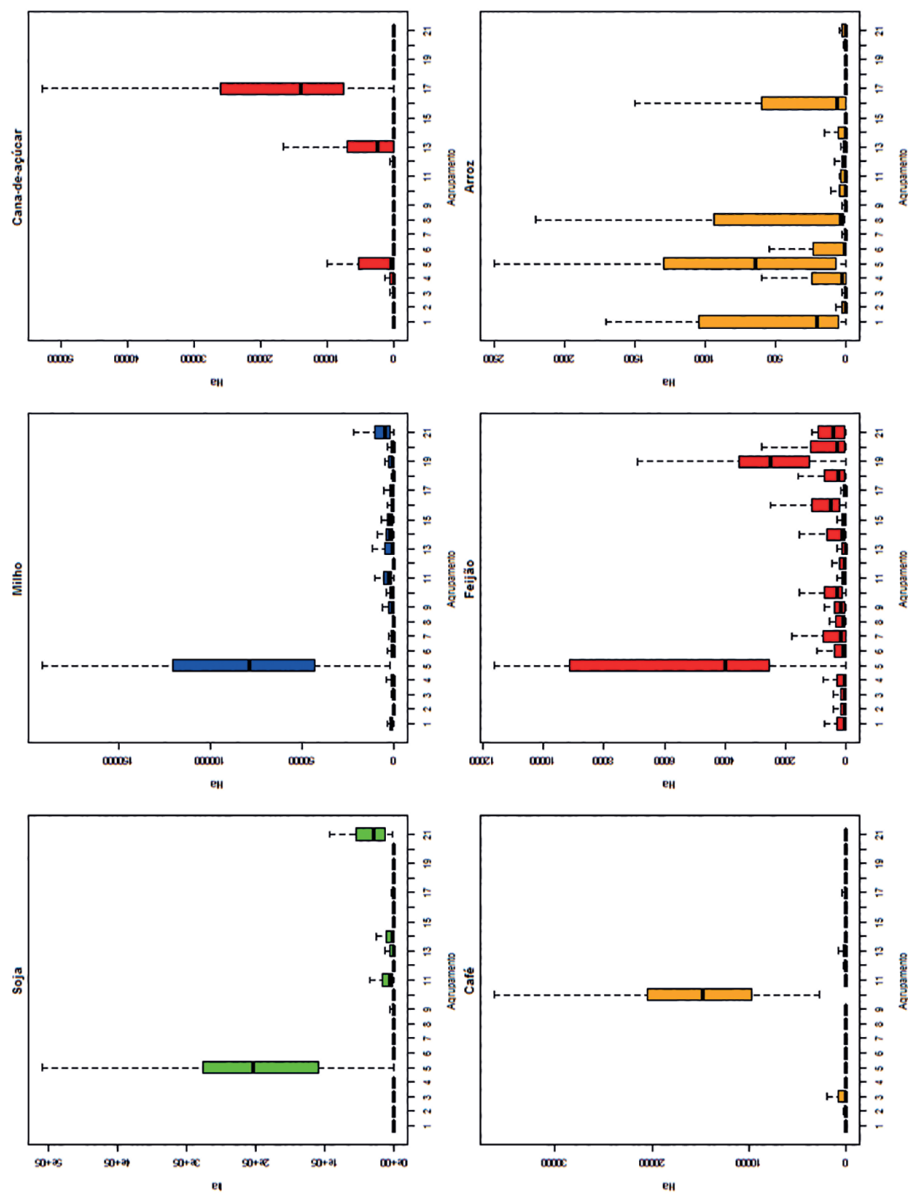
IBGE – INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **Produção agrícola municipal**. Rio de Janeiro: IBGE, 2012. v. 39.

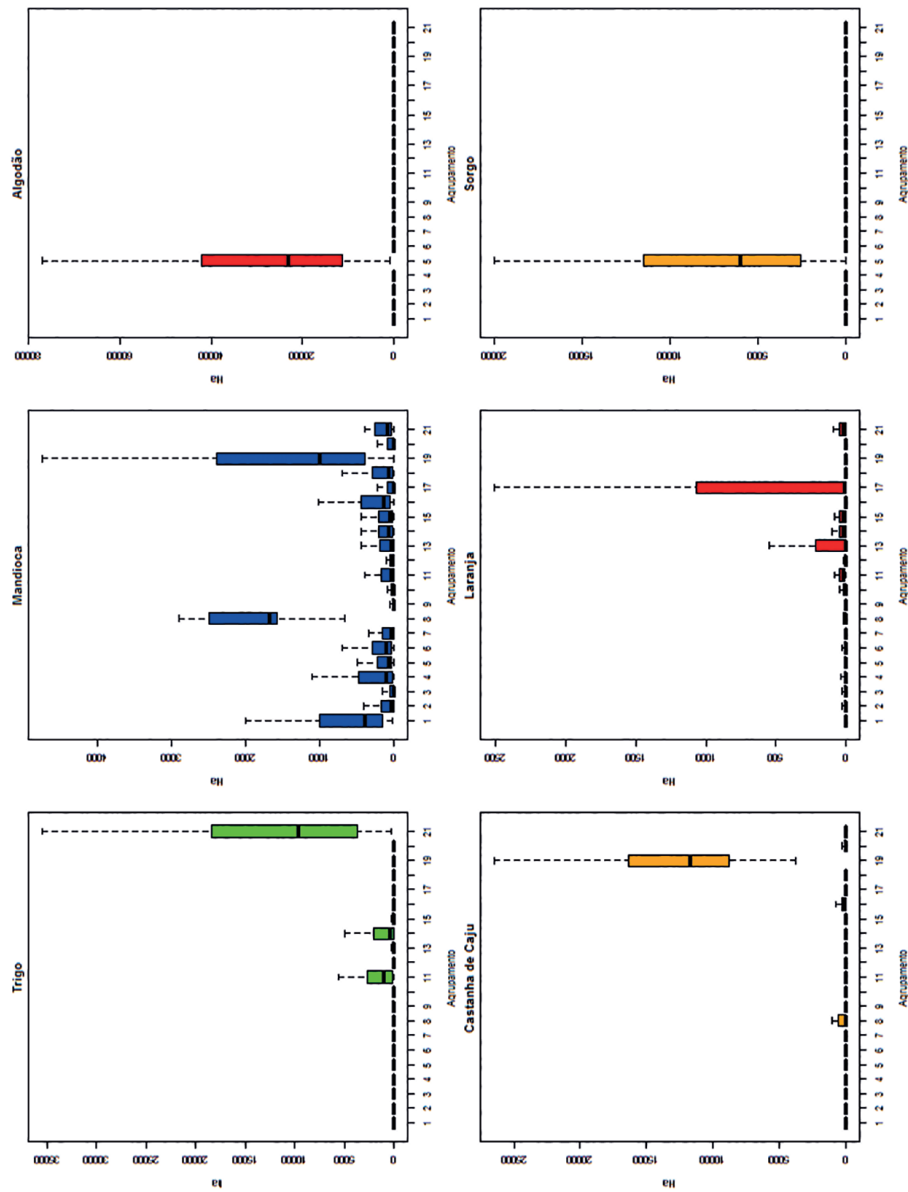
_____. **Produção da pecuária municipal**. Rio de Janeiro: IBGE, 2012. v. 40.

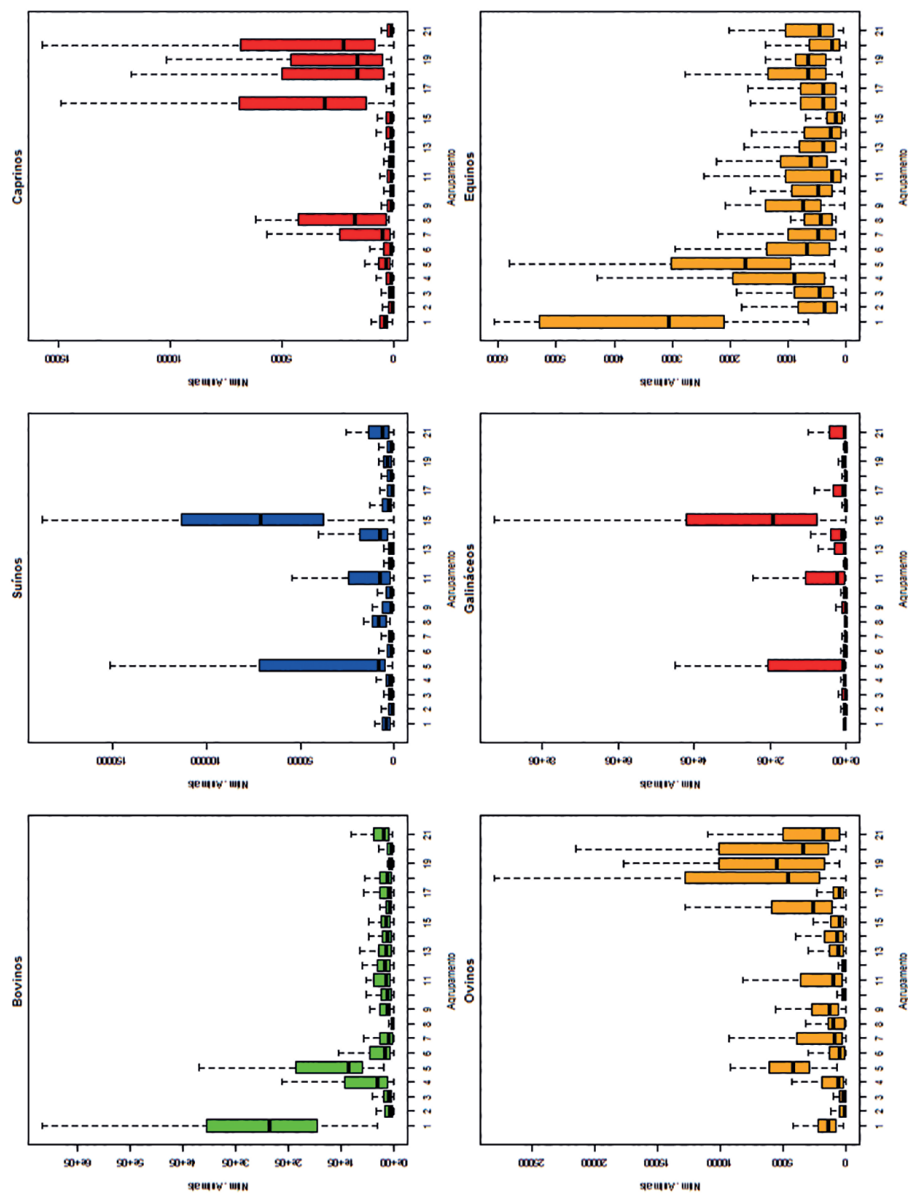
APÊNDICE A

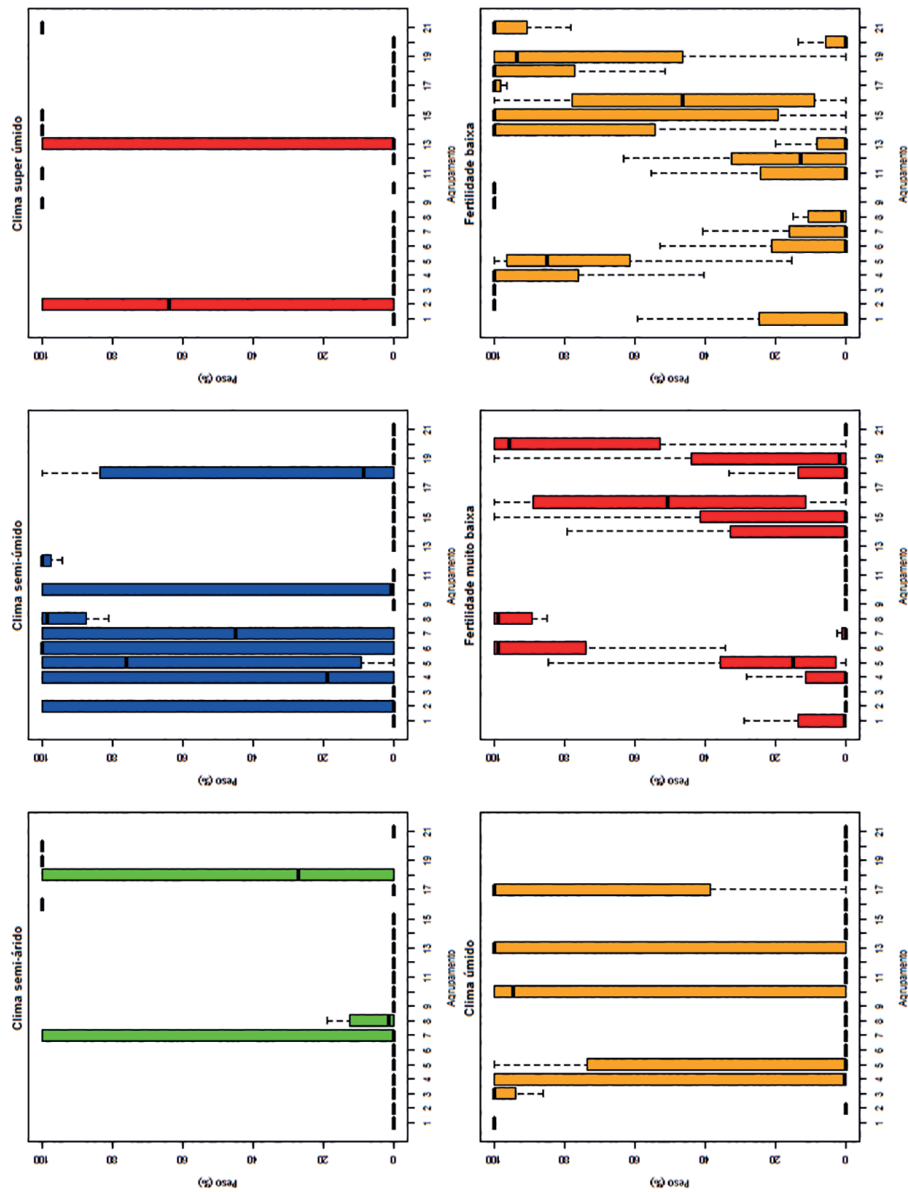
GRÁFICO A.1
Box-plots k-means

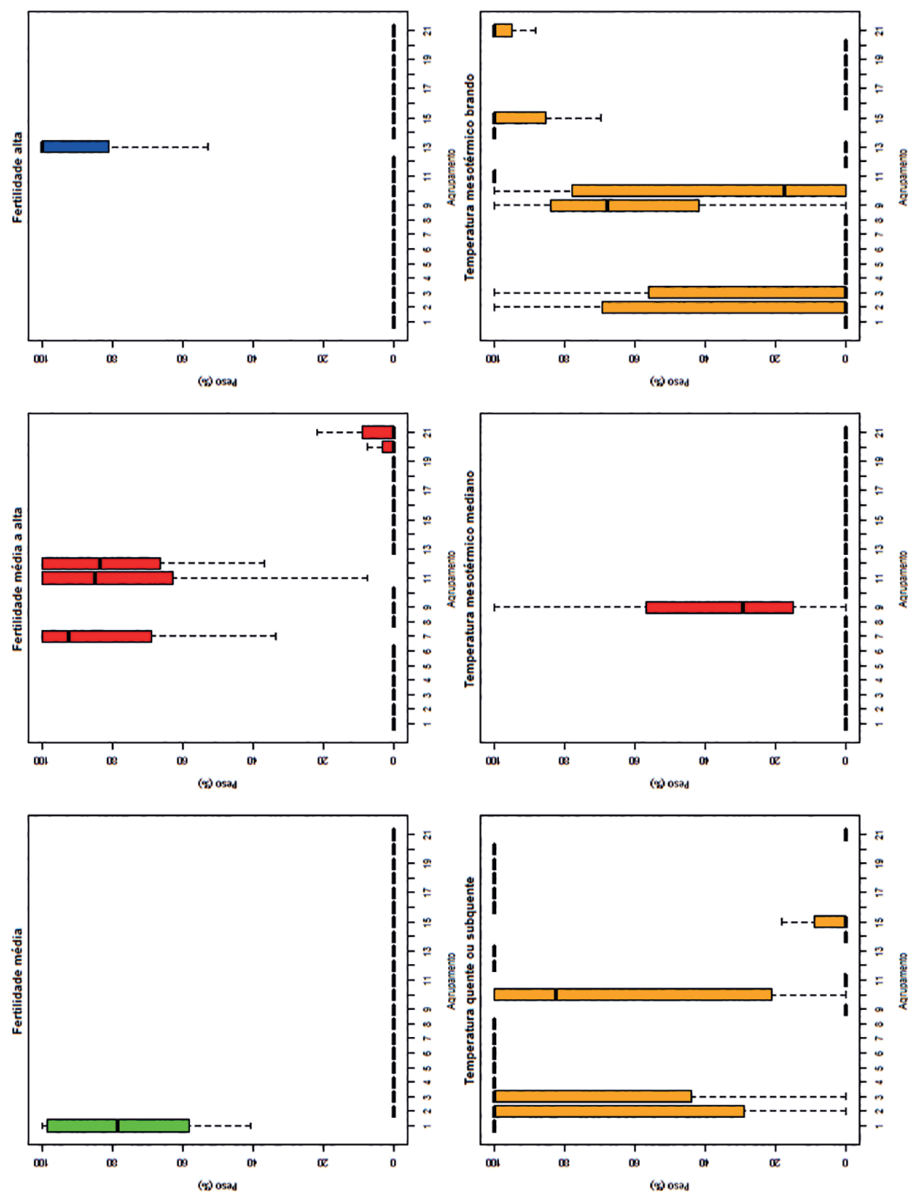








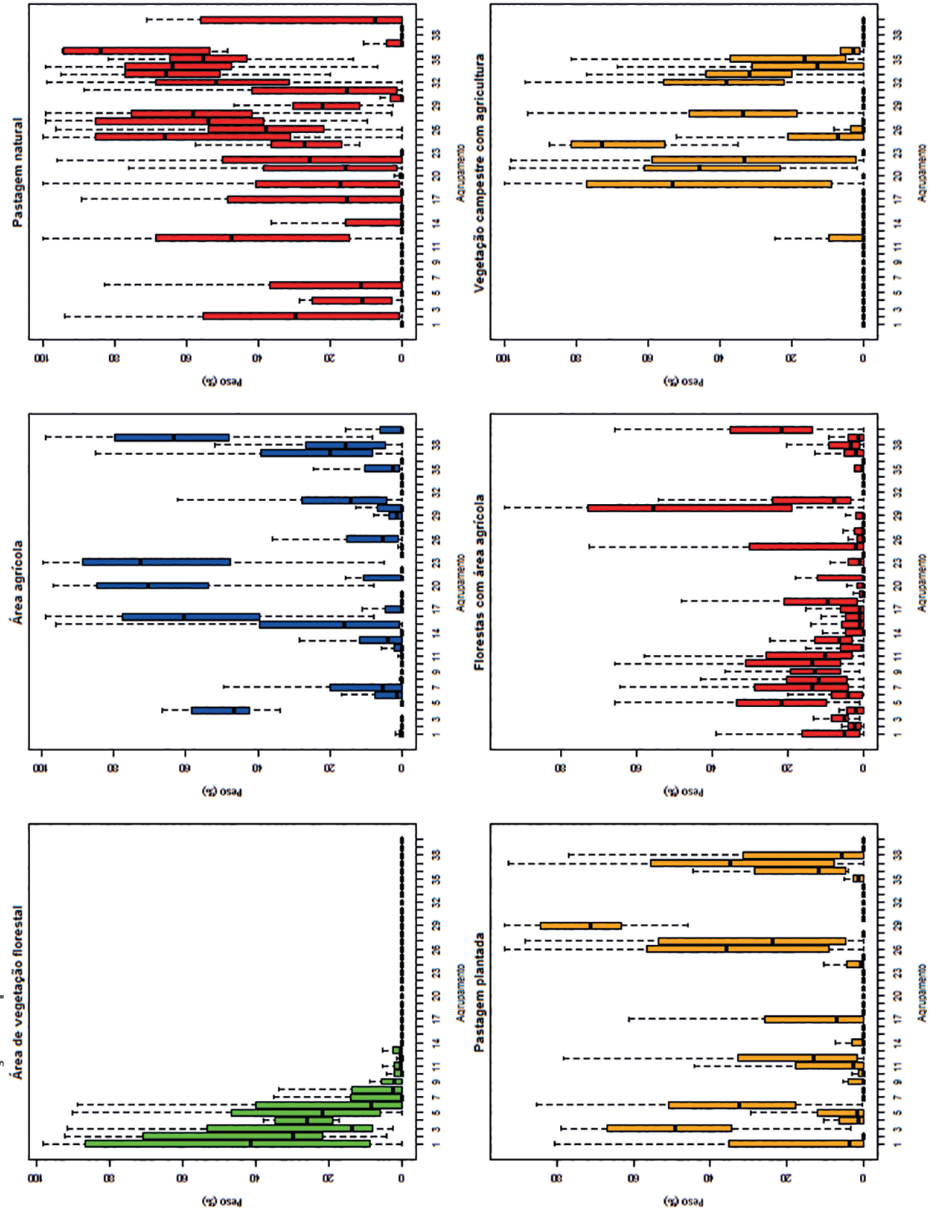


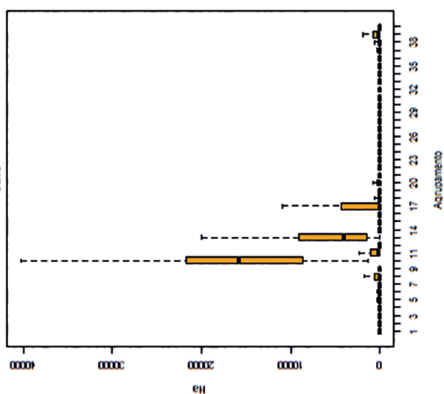
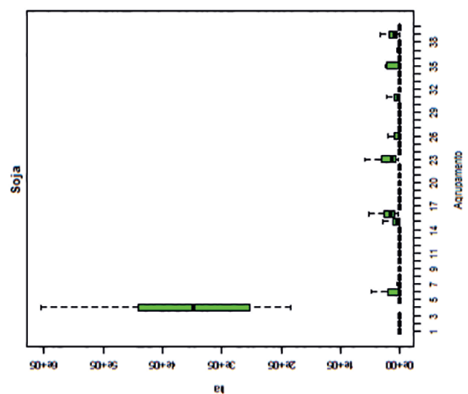
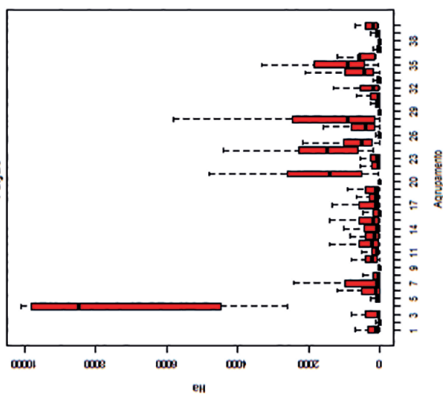
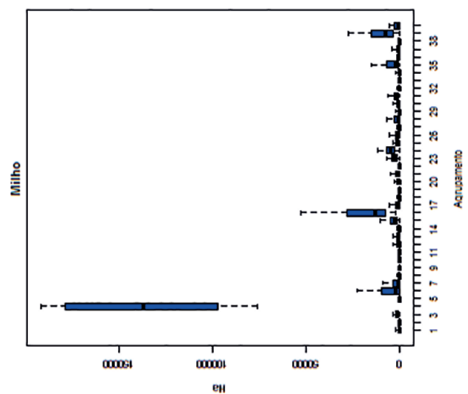
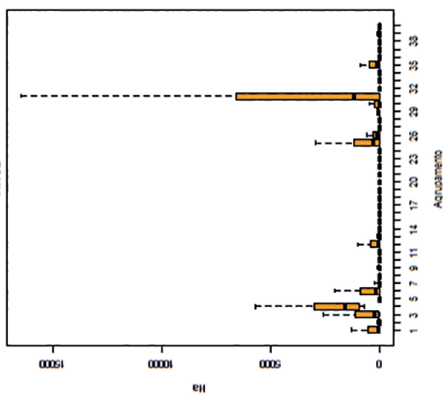
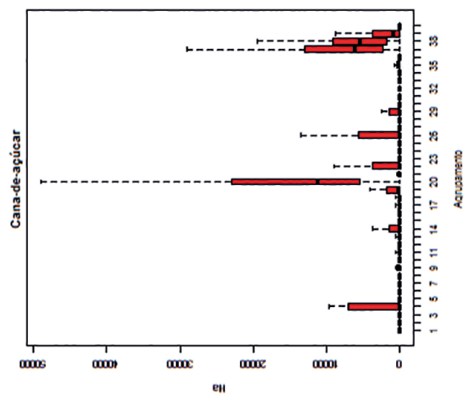


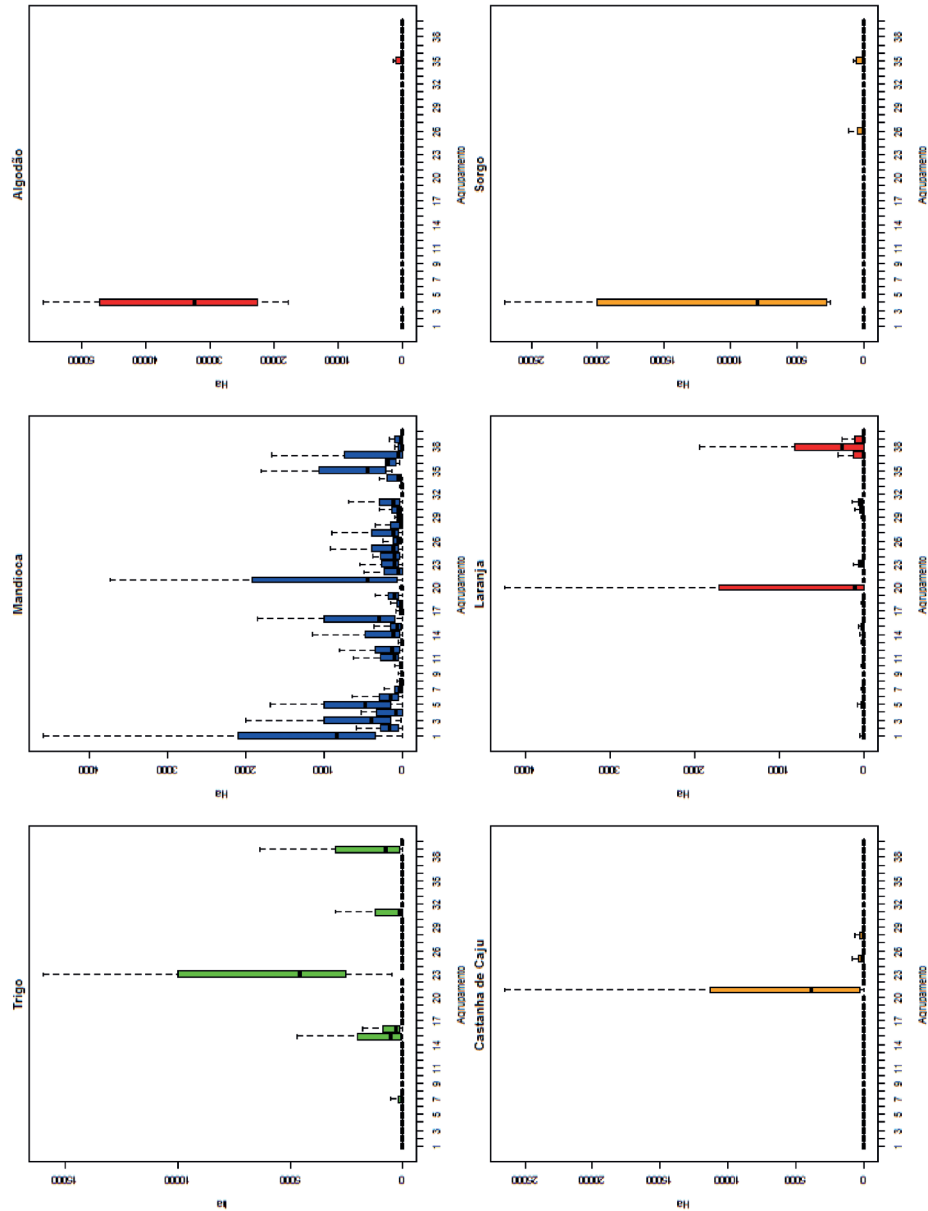
Elaboração dos autores.
 Obs.: Figura reproduzida em baixa resolução e cujos leiaute e textos não puderam ser padronizados e revisados em virtude das condições técnicas dos originais (nota do Editorial).

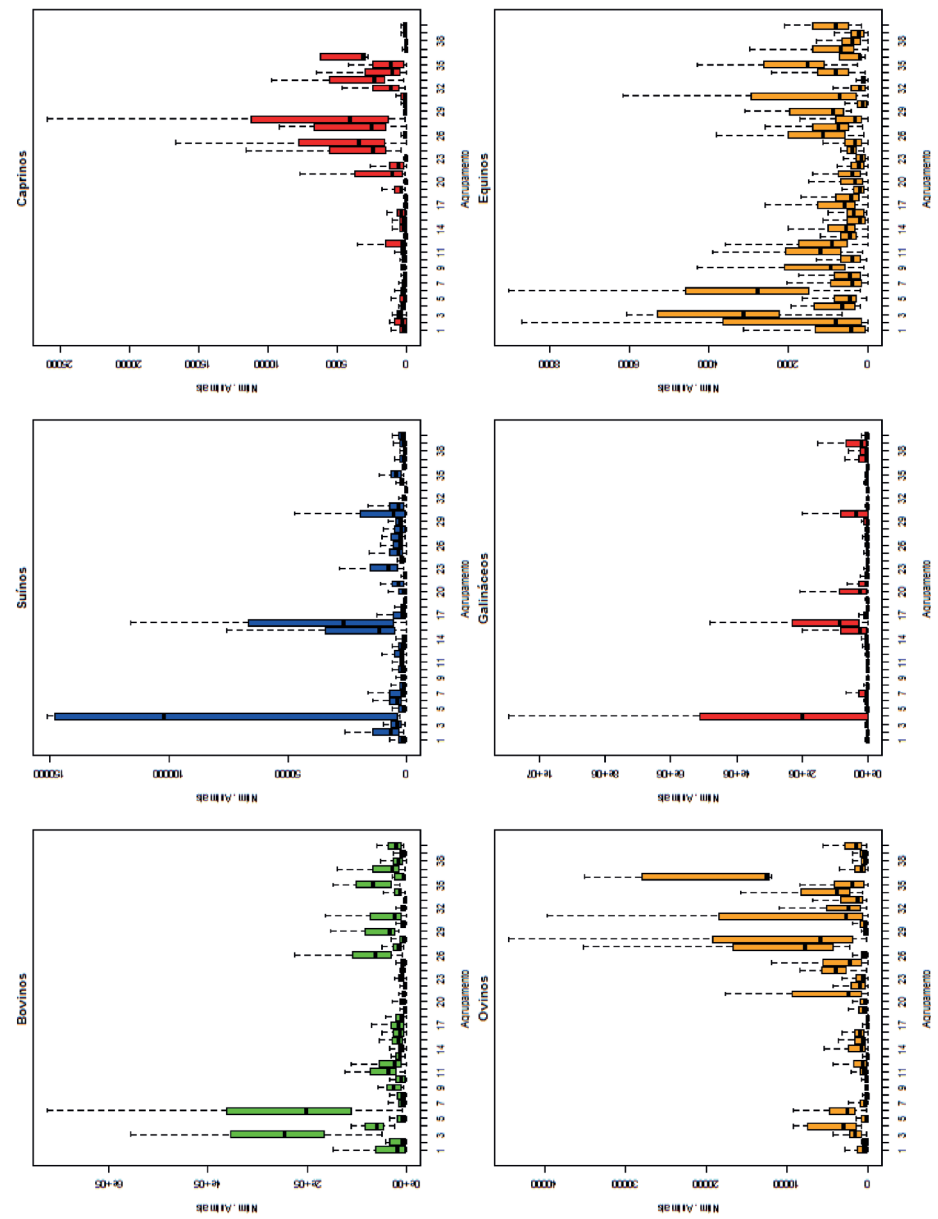
APÊNDICE B

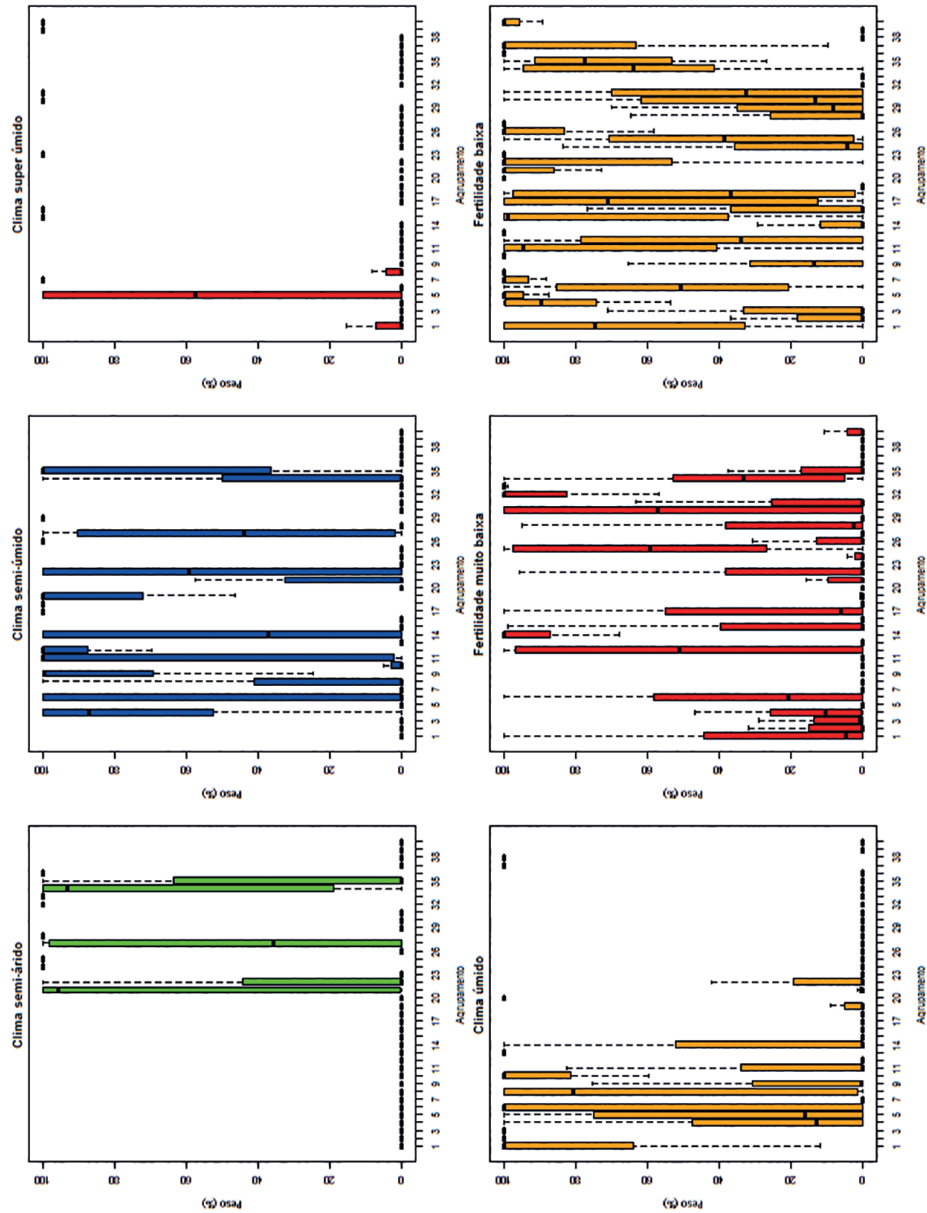
GRÁFICO B.1
Box-plots clusterização espacial
Área de vegetação florestal

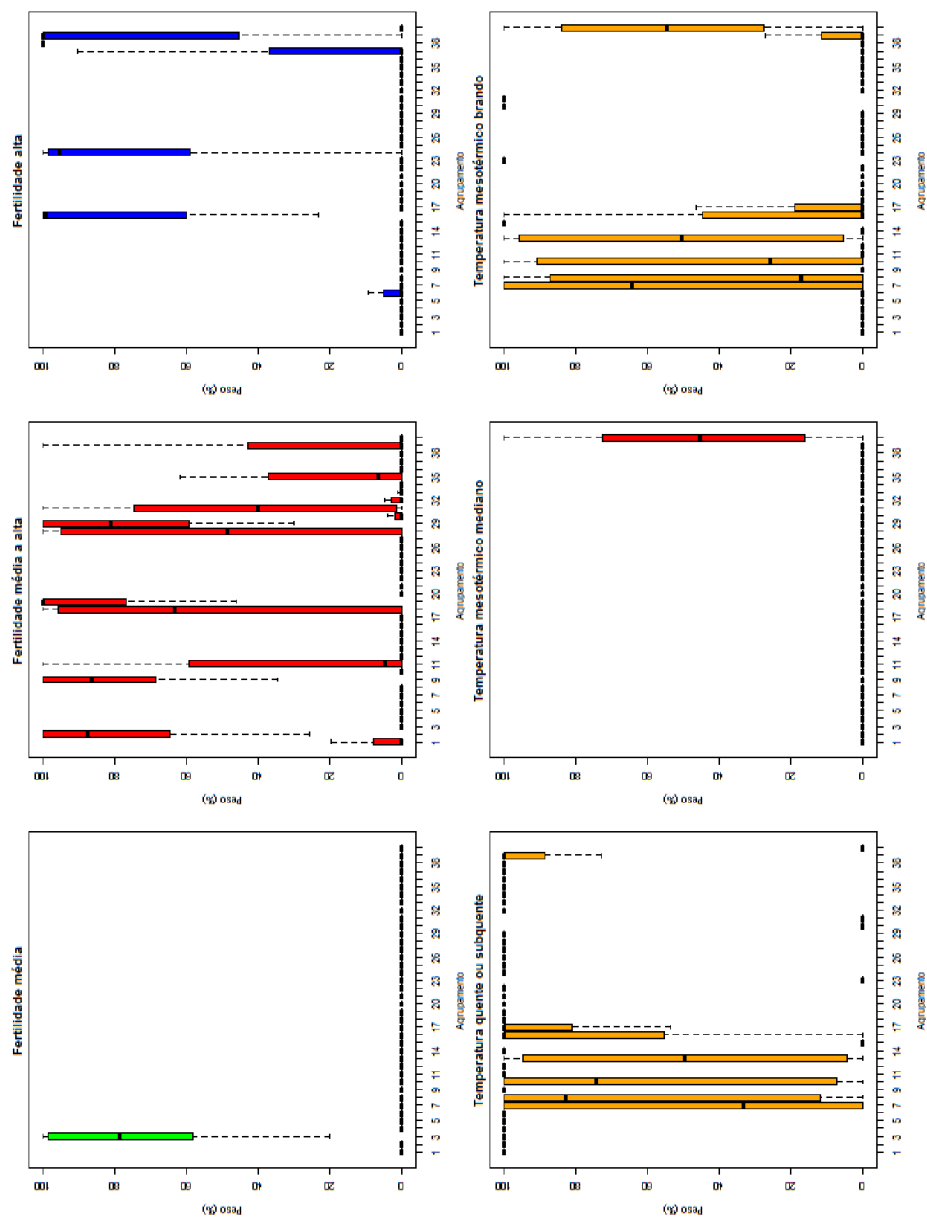












Elaboração dos autores.
Obs.: Figura reproduzida em baixa resolução e cujos leiaute e textos não puderam ser padronizados e revisados em virtude das condições técnicas dos originais (nota do Editorial).

APÊNDICE C

CATEGORIAS DAS VARIÁVEIS

QUADRO C.1
Características físicas

Valor	Descrição
Bom	Boa capacidade de armazenamento de água e boa aeração do solo, resultando em um ambiente favorável para agricultura.
Regular	Moderada capacidade de armazenamento de água e moderada aeração do solo, resultando em um ambiente regular para agricultura.
Ruim	Desfavorável capacidade de armazenamento de água e aeração do solo, resultando em um ambiente ruim para agricultura.

Elaboração dos autores.

QUADRO C.2
Fertilidade

Valor	Descrição
Alto	Saturação por bases superior a 50%, resultando em alta reserva de nutrientes para as plantas.
Médio alto	Entre média e alta reserva de nutrientes relacionados ao desenvolvimento das plantas.
Médio	Saturação por bases e por alumínio menor que 50%, resultando em média reserva de nutrientes para as plantas.
Médio baixo	Entre baixa e média reserva de nutrientes relacionados ao desenvolvimento das plantas.
Baixa	Saturação por bases menor que 50% e por alumínio maior que 50%, resultando em baixa reserva de nutrientes para as plantas.

Elaboração dos autores.

QUADRO C.3
Limitações do solo

Valor
Alta salinidade, profundidade reduzida, presença de pedregosidade, rochosidade e textura arenosa.
Baixo nível de nutrientes e excesso de alumínio.
Baixo nível de nutrientes, excesso de alumínio e textura grosseira.
Declives acentuados.
Declives acentuados, pouca profundidade e textura grosseira.
Declives acentuados, drenagem restrita e excesso de alumínio.
Excesso de sódio, drenagem restrita e risco de inundação.
Disponibilidade de nutrientes entre média e baixa.
Praticamente sem limitações.
Impedimento de drenagem e risco de inundação.

Elaboração dos autores.

QUADRO C.4

Seca

Valor
Entre um e dois meses de seca
Três meses de seca
Entre quatro e cinco meses de seca
Seis meses de seca
Entre sete e oito meses de seca
Entre nove e dez meses de seca
Onze meses de seca
Sem seca
Subseca

Elaboração dos autores.

QUADRO C.5

Tipos de solo

Valor	Descrição
Bom	Alta fertilidade natural, entre bem a moderadamente drenado, profundo e boa capacidade de armazenamento de água.
Bom a regular	Baixa fertilidade natural, bem drenado, muito profundo e boa capacidade de armazenar água.
Regular a bom	Média e baixa fertilidade natural, entre bem a moderadamente drenado, muito profundo e profundo e boa capacidade de armazenar água.
Regular	Média e baixa fertilidade natural, boa e moderada drenagem, muito profundo e profundo e boa capacidade de armazenar água.
Regular a restrito	Baixa e média fertilidade natural, boa drenagem, profundo, entre moderada a boa capacidade de armazenagem de água.
Restrito	Baixa fertilidade natural, de imperfeita a moderada drenagem, muito profundo e profundo, muito boa e boa capacidade de armazenamento de água.
Regular a desfavorável	Entre baixa e alta fertilidade natural, de moderada a imperfeita drenagem, entre muito profundo e raso, e de alta a baixa capacidade de armazenamento de água.
Desfavorável	Rocha exposta.

Elaboração dos autores.

QUADRO C.6

Temperatura

Valor	Descrição
Quente	Temperatura média maior que 18°C em todos os meses.
Subquente	Temperatura média entre 15°C e 18°C por pelo menos um mês.
Mesotérmico brando	Temperatura média entre 10°C e 15°C.
Mesotérmico mediano	Temperatura média menor que 10°C.

Elaboração dos autores.

**QUADRO C.7
Topologia**

Valor	Descrição
Plano e suavemente ondulado	Superfície predominantemente plana com suaves ondulações.
Plano a ondulado	Superfície predominantemente plana contendo ondulações mais íngremes.
Forte ondulado	Superfície contendo predominantemente ondulações acentuadas.
Ondulado a montanhoso	Superfície contendo ondulações e montanhas.
Montanhoso a escarpado	Superfície contendo montanhas e escarpas.

Elaboração dos autores.

**QUADRO C.8
Tipos de clima**

Valor	Descrição
Superúmido	Sem meses secos.
Úmido	De um a três meses secos.
Semiúmido	De quatro a cinco meses secos.
Semiárido	De seis a oito meses secos.
Árido	De nove a onze meses secos.

Elaboração dos autores.

**QUADRO C.9
Uso do solo**

Valor
Área agrícola
Área artificial
Área descoberta
Corpo d'água continental
Mosaico de agropecuária com remanescentes florestais
Mosaico de vegetação campestre com áreas agrícolas
Mosaico de vegetação florestal com áreas agrícolas
Pastagem natural
Pastagem plantada
Silvicultura
Vegetação campestre
Vegetação campestre alagada
Vegetação florestal
Outros

Elaboração dos autores.

VALORES MÉDIOS DAS PRINCIPAIS VARIÁVEIS, POR CLUSTER

TABELA D.1
Valor médio das principais culturas (em ha) e criações (em número de animais) para o método hierárquico não espacial com 21 clusters

Variáveis/cluster	1	2	3	4	5	6	7	8	9	10	11
Arroz	319,93	2.501,96	879,43	288,82	115,38	780,44	421,24	522,05	75,34	762,07	2.279,50
Café	503,66	4,03	66,71	1.152,67	336,51	486,80	84,75	230,64	504,77	146,90	64,49
Feijão	400,09	48,49	207,21	242,92	181,18	4.869,44	440,32	306,81	847,93	519,38	959,45
Mandioca	1.295,96	501,50	650,93	532,30	101,17	385,85	215,61	382,47	78,34	166,46	684,88
Soja	1.323,70	2.946,79	417,86	2.764,47	427,07	117,410,75	2.969,35	1.297,69	3.972,54	8.200,40	1.201,85
Cana-de-açúcar	385,47	21,76	17,25	3.074,95	824,43	9.422,48	290,95	989,67	1.310,40	1.695,12	211,27
Milho	1.147,48	603,84	1.610,71	2.071,91	798,24	53.957,31	2.850,96	1.093,14	3.133,12	3.031,73	1.330,71
Sorgo	21,79	38,16	0,00	60,99	1,05	5.910,64	1,49	35,31	214,75	269,08	52,07
Algodão	58,38	0,00	0,00	83,91	0,14	14.878,00	0,00	33,39	129,34	407,05	12,31
Castanha-de-caju	26,16	0,51	1,64	3,94	0,27	0,00	0,00	8,70	3,08	51,36	1.461,43
Laranja	27,11	31,78	4,00	106,82	77,80	43,40	50,76	18,33	263,89	28,94	80,30
Trigo	0,00	196,15	0,00	21,17	96,26	185,40	359,16	1,80	331,43	5,74	164,77
Bovinos	46.184,82	28.620,16	282.991,29	134.267,40	15.717,41	11.0761,83	19.473,44	39.796,00	18.677,38	60.898,14	23.077,80
Equinos	992,18	1.358,62	3.412,46	1.985,59	681,97	2.325,40	681,17	996,28	863,10	1.269,71	1.114,22
Suínos	2.997,95	6.719,78	4.712,04	4.499,34	5.243,00	50.271,48	30.747,86	3.304,50	4.613,59	4.106,48	4.247,06
Galináceos	43.544,44	79.224,63	41.488,21	147.936,03	253.786,18	1.259.899,75	756.826,27	90.416,99	186.425,78	168.075,42	208.428,04
Caprinos	561,08	438,38	503,75	258,51	270,80	600,08	352,88	516,70	202,68	431,41	8.301,31
Ovinos	1.071,29	3.308,40	1.955,79	2.071,47	558,22	3.594,43	1.397,41	1.056,07	1.203,04	1.092,14	14.747,98
Número de áreas mínimas comparáveis (AMCs)	291	68	28	458	642	75	276	505	180	435	355

(Continua)

(Continuação)

Variáveis/cluster	12	13	14	15	16	17	18	19	20	21
Arroz	37,59	110,21	42,35	17,96	203,42	15,67	18,19	702,69	41,69	84,17
Café	13.650,85	729,82	327,64	649,78	71,97	332,92	38,53	82,89	527,81	0,33
Feijão	565,77	192,35	678,26	153,47	1.396,12	86,57	680,12	884,69	153,14	1.028,56
Mandioca	65,06	46,62	134,71	122,88	175,79	215,53	222,18	388,59	370,75	78,59
Soja	581,54	282,80	1.610,06	286,71	16.796,09	1.421,45	7,14	488,31	14.043,42	241,33
Cana-de-açúcar	486,99	294,34	389,54	5.448,08	89,64	18.997,23	196,47	192,94	3.133,54	18,70
Milho	1.735,91	591,43	1.837,52	741,65	4.830,02	1.320,75	635,24	1.752,55	12.954,47	1.591,32
Sorgo	42,42	0,28	8,38	24,31	8,41	77,08	80,28	23,42	0,00	6,84
Algodão	22,49	0,00	7,13	6,08	0,30	7,57	57,22	20,91	7,78	3,81
Castanha-de-caju	0,00	0,00	32,19	0,00	0,00	5,93	156,48	405,17	0,00	193,57
Laranja	37,66	11,27	16,52	423,13	51,67	2.168,41	6,22	5,84	185,55	2,41
Trigo	37,41	0,00	283,93	3,10	4.530,28	2,03	0,00	0,00	1.231,97	23,72
Bovinos	23.978,65	28.521,02	20.479,08	36.529,19	22.713,42	20.987,71	17.214,14	11.622,81	11.784,81	16.150,93
Equinos	823,60	1.016,85	693,69	837,45	685,84	736,45	882,74	657,82	368,56	685,36
Suínos	6.076,31	2.915,19	6.050,20	1.771,52	17.735,00	2.734,15	31.14,39	5.228,30	9.823,39	2.917,56
Galináceos	179.376,42	50.001,55	211.514,91	149.308,04	493.480,37	264.383,31	88.217,48	69.285,93	631.438,24	39.504,62
Caprinos	203,35	214,91	2.182,54	148,60	337,59	130,01	4.513,42	5.252,38	217,92	5.309,99
Ovinos	353,46	315,62	3.369,00	839,44	2.703,52	1.120,18	8.761,32	4.229,64	874,61	11.671,32
Número de AMCs	182	100	387	210	299	182	161	327	118	196

Elaboração dos autores.

TABELA D.2
Valor médio das principais culturas (em ha) e criações (em número de animais) para o método *k-means* com 21 clusters

Variáveis/clusters	1	2	3	4	5	6	7	8	9	10	11
Arroz	871,17	151,36	140,23	408,13	1.056,39	561,64	55,70	527,25	14,42	42,11	4.369,69
Café	64,41	461,62	1.204,32	574,99	1.086,06	241,79	248,80	0,00	14,40	16.381,87	163,52
Feijão	220,07	210,91	161,12	430,46	8.809,03	449,41	767,64	185,38	455,76	587,81	161,65
Mandioca	669,62	201,36	196,70	631,38	610,77	351,15	246,12	2.227,75	32,20	61,45	153,53
Soja	403,45	124,54	8,90	4.940,29	212.381,94	3.135,14	89,91	0,00	2.105,14	320,11	11.574,80
Cana-de-açúcar	16,66	121,93	917,65	1.539,77	5.168,97	807,67	444,05	1,50	173,54	541,67	318,35
Milho	1.989,79	572,27	510,35	2.520,81	91.886,90	1.740,60	1.152,73	526,25	2.260,64	1.785,04	4.574,51
Sorgo	0,00	2,26	0,21	159,49	8.506,39	113,66	97,92	0,00	0,70	42,23	134,77
Algodão	0,00	0,53	0,25	221,48	32.971,42	143,39	15,37	0,00	0,00	10,07	0,46
Castanha-de-caju	1,59	0,18	0,58	30,33	0,00	5,64	100,16	500,25	0,00	0,00	0,00
Laranja	3,86	27,42	86,34	59,25	30,00	109,58	8,08	7,00	14,54	46,89	63,42
Trigo	0,00	30,39	0,17	16,28	281,48	40,01	1,72	0,00	226,40	37,20	2.188,07
Bovinos	347.155,72	15.964,18	22.681,88	87.276,38	135.177,29	44.185,98	22.178,97	4.369,38	23.201,68	21.876,53	64.765,24
Equinos	4.011,48	668,74	744,06	1.568,32	2.361,74	1.126,47	845,74	495,13	1.137,36	817,71	2.171,76
Suínos	5.553,90	2.500,14	3.738,97	4.797,13	71.489,58	3.936,65	2.600,76	8.246,50	6.972,78	4.129,98	17.483,01
Galináceos	49.584,62	141.535,87	218.649,21	136.651,14	1.730.158,61	126.672,44	77.626,94	28.413,00	190.177,50	105.993,01	639.563,36
Caprinos	593,07	237,18	236,80	501,64	610,68	456,03	2.886,82	2.396,38	220,14	218,17	348,42
Ovinos	2.195,83	435,11	500,31	1.571,68	4.657,23	1.502,04	3.763,91	1.044,25	2.385,64	362,91	26.818,32
Número de AMCs	29	470	355	959	31	778	352	8	50	139	108

(Continua)

(Continuação)

Variáveis/clusters	12	13	14	15	16	17	18	19	20	21
Arroz	31,23	35,93	710,04	10,54	682,67	66,17	1.712,43	11,89	45,87	101,52
Café	520,75	616,18	45,53	151,59	73,51	418,17	37,37	1,11	2,27	41,29
Feijão	188,82	194,50	974,58	193,65	952,85	193,78	693,82	2.775,26	1.130,15	1.324,29
Mandioca	43,40	232,86	171,63	250,89	432,41	153,14	396,94	1.613,89	118,81	295,50
Soja	37,27	4.714,65	7.300,61	3.846,84	853,92	6.287,03	307,97	0,00	3,60	44.883,02
Cana-de-açúcar	312,13	5.109,12	100,57	586,66	244,79	18.183,63	216,48	70,19	23,98	115,06
Milho	440,84	4.579,79	3.605,77	4.725,54	1.638,23	4.611,87	658,20	1.899,93	1.622,20	7.455,67
Sorgo	0,32	41,05	10,17	69,97	25,24	434,02	76,32	12,96	17,20	12,71
Algodão	0,00	8,62	8,20	4,25	47,48	121,50	55,70	3,89	5,34	1,04
Castanha-de-caju	0,00	4,38	0,00	0,00	425,59	4,30	138,90	14.517,78	394,56	0,00
Laranja	10,42	334,31	54,44	99,47	5,36	1.523,70	7,30	1,07	2,17	35,69
Trigo	0,00	417,97	1.528,44	297,39	0,00	73,50	9,15	0,00	0,00	13.578,50
Bovinos	25.242,50	30.998,64	18.738,77	20.094,78	12.001,32	25.494,41	24.984,20	7.190,74	10.200,40	34.963,31
Equinos	919,80	679,82	657,77	378,32	624,72	708,58	1.124,71	695,85	542,72	898,50
Suínos	2.918,48	4.139,17	14.919,01	10.2372,59	5.234,65	3.684,16	3.035,55	4.000,63	2.954,00	15.505,63
Galináceos	44.369,09	286.520,07	340.346,07	2.878.977,11	63.959,67	412.241,09	80.171,70	213.378,26	37.369,07	611.134,10
Caprinos	186,36	327,54	334,70	374,99	5.259,13	123,82	5.807,09	2.983,00	10.099,68	262,48
Ovinos	267,48	1.103,55	1.773,56	967,15	4.405,48	1.050,15	11.356,15	6.918,33	11.249,79	6.026,04
Número de AMCs	88	353	507	79	350	262	164	27	318	48

Elaboração dos autores.

TABELA D.3
Valor médio das principais culturas (em ha) e criações (em número de animais) para o método hierárquico espacial com quarenta clusters

Variáveis/clusters	1	2	3	4	5	6	7	8	9	10
Arroz	738,91	256,88	909,17	2.195,00	0,00	1.137,44	278,10	60,98	41,70	33,59
Café	91,20	1,25	181,66	1,25	194,14	1.325,04	269,75	1.297,07	24,15	16.796,03
Feijão	304,47	27,06	236,83	10.688,63	111,76	798,00	1.139,29	193,82	31,07	323,66
Mandioca	1.946,99	227,50	669,62	173,13	769,81	327,98	137,03	74,98	18,63	82,39
Soja	426,02	0,00	421,79	359.435,38	0,00	23.260,58	3.574,13	13,96	0,00	0,00
Cana-de-açúcar	74,92	3,31	16,66	5.024,13	50,41	1.516,56	805,90	471,30	328,48	481,71
Milho	1.145,58	45,94	2.029,10	156.483,88	97,96	9.784,00	2.862,72	494,75	101,33	536,69
Sorgo	0,00	0,00	0,00	11.387,50	0,00	522,89	15,52	0,30	0,00	0,00
Algodão	0,00	0,00	0,00	41.061,00	0,00	2.320,98	39,91	0,00	0,00	0,00
Castanha-de-caju	20,72	0,00	1,59	0,00	2,00	3,29	0,00	0,00	0,00	0,00
Laranja	71,78	11,56	3,41	0,00	145,13	7,36	189,25	32,79	7,07	13,00
Trigo	0,00	0,00	0,00	0,00	0,00	9,93	770,51	7,87	0,00	0,00
Bovinos	54.719,72	20.286,25	356.357,48	64.796,63	13.383,40	271.851,56	13.366,73	13.559,83	30.980,44	21.011,81
Equinos	930,49	2.358,63	4.123,86	849,25	658,07	3.619,17	735,76	696,29	1.417,59	748,05
Suínos	2.977,35	10.705,38	5.688,03	118.981,00	2.440,34	8.100,89	9.385,05	4.382,79	1.504,67	2.813,73
Galináceos	72.411,78	11.362,38	51.653,79	3.153.945,25	116.658,53	158.544,00	316.666,18	237.624,81	104.592,93	49.231,39
Caprinos	417,40	522,31	610,34	320,63	420,87	362,05	266,91	161,03	302,26	234,64
Ovinos	1.041,70	693,81	2.254,48	4.247,50	606,90	3.940,31	1.084,21	305,59	339,22	372,66
Número de AMCs	250	16	29	8	70	219	328	394	27	59

(Continua)

(Continuação)

Variáveis/clusters	11	12	13	14	15	16	17	18	19	20
Arroz	20,10	592,88	42,05	29,71	20,93	23,78	19,75	28,13	0,26	12,34
Café	1.546,79	501,44	5.987,38	6,35	51,74	174,11	4.443,10	890,47	0,00	460,24
Feijão	169,98	837,95	274,07	570,60	760,07	196,58	559,15	218,11	305,47	100,17
Mandioca	230,70	347,88	42,02	429,70	134,68	703,31	40,97	55,81	216,93	70,74
Soja	0,00	4.164,30	62,06	0,00	8.401,44	21.166,81	1.787,48	2,37	0,00	1.137,52
Cana-de-açúcar	892,73	366,25	1.015,02	2.158,08	67,52	210,28	1.134,52	219,81	1.715,89	15.926,39
Milho	234,15	1.918,46	1.314,15	1.113,69	4.327,28	20.698,44	2.594,96	541,10	235,16	1.330,89
Sorgo	0,28	324,27	1,17	2,34	1,40	0,00	161,64	0,70	0,00	97,09
Algodão	0,00	214,08	0,00	0,25	0,63	0,31	22,66	0,00	0,51	4,97
Castanha-de-caju	0,00	18,29	0,00	107,62	0,00	0,00	0,00	0,00	28,04	0,00
Laranja	9,73	11,87	79,31	480,26	36,38	1,11	32,67	11,19	13,98	1.865,33
Trigo	0,00	20,13	11,11	0,00	1.759,34	750,97	76,56	0,00	0,00	4,72
Bovinos	53.495,96	43.109,19	16.646,52	12.717,93	22.081,98	16.767,00	27.898,96	16.979,69	4.979,05	11.604,88
Equinos	1.631,44	1.355,08	525,10	799,42	484,79	353,22	944,91	625,02	280,18	495,51
Suínos	2.574,52	4.272,84	2.328,11	1.715,81	28.118,02	56.092,36	11.638,34	2.624,31	501,74	3.239,82
Galináceos	21.806,02	89.011,09	86.625,75	136.785,03	633.790,75	1.519.925,11	351.380,76	23.179,03	62.781,16	627.794,52
Caprinos	305,58	1.981,43	73,22	709,39	364,23	590,86	82,75	102,21	670,07	91,83
Ovinos	887,41	1.944,98	215,67	2.482,69	1.382,09	1.204,86	251,56	128,42	744,82	687,62
Número de AMICS	116	581	81	239	334	36	209	118	57	156

(Continua)

(Continuação)

Variáveis/clusters	21	22	23	24	25	26	27	28	29	30
Arroz	62,59	0,05	8,27	0,00	927,42	242,60	0,00	74,25	92,29	1.302,96
Café	205,68	21,40	0,00	154,67	21,62	22,24	0,11	2,74	101,65	0,00
Feijão	1.967,86	220,87	237,69	1.644,00	860,02	252,23	663,03	1.880,12	71,32	123,09
Mandioca	1.244,48	200,16	211,66	293,33	428,90	97,24	370,66	222,49	97,91	140,48
Soja	0,00	0,00	28.164,53	0,00	814,87	10.629,56	0,00	4,29	460,88	1.099,20
Cana-de-açúcar	352,77	2.305,53	80,93	58,80	215,79	4.764,98	0,09	77,25	1.585,94	38,25
Milho	1.783,73	263,23	3.570,97	5.074,00	1.476,94	4.859,27	649,31	2.186,89	786,47	1.279,97
Sorgo	7,95	0,00	7,36	79,33	5,16	1.240,21	0,00	14,81	280,29	0,45
Algodão	0,00	2,96	0,00	29,33	2,21	257,39	0,00	13,61	0,00	0,00
Castanha-de-caju	6.961,27	167,31	0,00	0,00	493,53	0,47	27,09	826,48	0,00	0,00
Laranja	5,52	3,78	51,04	0,93	6,35	46,91	12,43	2,84	39,65	58,05
Trigo	0,00	0,00	8.645,81	0,00	0,00	4,35	0,00	0,00	0,00	141,17
Bovinos	6.500,50	5.175,84	20.041,78	8.367,53	9.204,93	91.897,91	21.225,17	11.107,22	56.006,62	9.754,71
Equinos	547,02	360,38	421,99	437,67	471,25	1.604,21	1.317,86	615,29	1.324,09	368,23
Suínos	4.349,27	1.043,51	14.754,57	2.252,27	6.041,98	10.950,03	6.010,26	3.660,81	4.317,32	14.193,91
Galináceos	312.291,05	115.987,09	57.553,18	17.374,27	54.714,46	206.301,86	101.259,83	54.687,33	271.226,03	733.682,39
Caprinos	2.310,36	1.114,12	114,19	4.269,53	5.977,86	139,48	5.492,60	13.689,09	90,35	161,03
Ovinos	5.088,43	1.648,46	3.150,78	4.375,80	4.296,97	757,16	13.771,09	15.829,35	295,65	978,31
Número de AMCs	44	171	74	15	252	160	35	267	34	158

(Continua)

(Continuação)

Variáveis/clusters	31	32	33	34	35	36	37	38	39	40
Arroz	6.609,84	37,04	0,00	0,00	474,53	0,00	128,23	4,28	51,11	10,75
Café	0,00	4,58	0,00	10,82	911,75	0,00	128,84	537,56	719,82	0,00
Feijão	211,96	616,66	86,92	834,97	1.912,00	534,00	100,74	33,20	136,89	287,81
Mandioca	247,17	67,86	17,72	493,25	1.100,94	345,00	519,98	87,43	150,93	11,06
Soja	8.195,79	0,00	0,00	0,00	36.426,47	0,00	6.585,49	109,95	12.565,88	1.372,22
Cana-de-açúcar	72,78	21,42	0,00	17,90	391,25	0,00	9.756,09	7.155,91	2.952,70	0,00
Milho	2.445,50	959,38	214,36	1.119,90	8.816,47	534,00	5.109,79	470,07	10.473,91	1.908,34
Sorgo	123,39	0,79	0,00	10,85	1.103,44	0,00	30,71	16,86	0,00	0,00
Algodão	0,00	2,45	0,08	4,52	12.024,56	0,00	29,31	4,15	6,63	0,00
Castanha-de-caju	0,00	237,86	29,26	66,11	0,94	0,00	0,00	0,00	0,00	0,00
Laranja	52,68	1,22	0,13	1,82	23,41	0,00	221,98	784,63	103,61	1,63
Trigo	1.055,54	0,00	0,00	0,00	0,00	0,00	66,39	0,00	2.362,54	104,69
Bovinos	70.618,42	8.615,50	3.117,44	18.289,72	68.672,91	13.808,60	58.332,96	19.609,35	10.143,61	31.035,16
Equinos	2.595,88	373,38	125,49	967,75	2.022,38	563,00	1.129,29	519,21	356,76	1.187,50
Suínos	8.090,42	1.787,72	440,28	2.553,56	4.842,16	1.292,00	3.974,67	1.572,89	3.191,00	5.740,09
Galináceos	76.353,35	28.661,20	9.643,05	78.055,30	64.917,66	18.940,00	306.777,41	271.741,82	509.387,62	120.916,31
Caprinos	421,17	2.109,48	3.857,74	3.460,08	1.830,63	6.200,00	118,71	61,90	138,99	153,91
Ovinos	26.426,55	4.107,23	2.288,64	6.591,89	2.900,72	19.940,00	1.585,60	662,11	826,64	2.747,19
Número de AMCs	137	146	39	61	32	5	246	150	90	32

Elaboração dos autores.

EDITORIAL

Coordenação

Cláudio Passos de Oliveira

Supervisão

Andrea Bossle de Abreu

Revisão

Carlos Eduardo Gonçalves de Melo

Elaine Oliveira Couto

Luciana Nogueira Duarte

Mariana Silva de Lima

Vivian Barros Volotão Santos

Editoração

Aeromilson Mesquita

Aline Cristine Torres da Silva Martins

Carlos Henrique Santos Vianna

Glaucia Soares Nascimento (estagiária)

Capa

Luís Cláudio Cardoso da Silva

Projeto Gráfico

Renato Rodrigues Bueno

The manuscripts in languages other than Portuguese published herein have not been proofread.

Livraria Ipea

SBS – Quadra 1 - Bloco J - Ed. BNDES, Térreo.

70076-900 – Brasília – DF

Fone: (61) 2026-5336

Correio eletrônico: livraria@ipea.gov.br

Missão do Ipea

Aprimorar as políticas públicas essenciais ao desenvolvimento brasileiro por meio da produção e disseminação de conhecimentos e da assessoria ao Estado nas suas decisões estratégicas.

ipea Instituto de Pesquisa
Econômica Aplicada

MINISTÉRIO DO
**PLANEJAMENTO,
DESENVOLVIMENTO E GESTÃO**

