Tema
Tendências em Matemática Aplicada e Computacional, **18**, N. 1 (2017), 69-84

www.scielo.br/tema
doi: 10.5540/tema.2017.018.01.0069

Clusterização Espacial e Não Espacial: Um Estudo Aplicado à Agropecuária Brasileira

M.G. PENA[1], G.C.C. MOREIRA[1], L.F.D. GUIMARÃES[1], C.R. LAURETO[1], P.H.M. ALBUQUERQUE[2*], A.X.Y. CARVALHO[1] e G.G. BASSO[1],

Recebido em 13 outubro 2016 / Aceito em 17 dezembro 2016

RESUMO. Este trabalho apresenta uma análise de clusterização de Áreas Mínimas Comparáveis (AMC's) 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-se, 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. Ainda, as metodologias discutidas permitem a incorporação de critérios tradicionais de escolha do número de *clusters*, tais como estatísticas *CCC*, *pseudo-F* e *pseudo-t*^{2}.

Palavras-chave: Clusterização espacial, algoritmos hierárquicos, *k-means*.

1 INTRODUÇÃO

Áreas territoriais extensas podem implicar em 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 as diversidades acima elencadas 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 ele se estende por quase todo o território nacional e é afetado com alterações nos padrões locacionais.

*Autor correspondente: Pedro Henrique Melo Albuquerque – E-mail: pedroa@unb.br
[1] Instituto de Pesquisa Econômica Aplicada – IPEA, 70076-900 Brasília, DF, Brasil.
E-mails: marina.pena@ipea.gov.br; guilhermechadud@gmail.com; luiz.guimaraes@ipea.gov.br; camilo.laureto@ipea.gov.br; alexandre.ywata@ipea.gov.br; gubasso@gmail.com
[2] Campus Universitário Darcy Ribeiro, Faculdade de Economia, Administração e Contabilidade (FACE), Universidade de Brasília – UnB, 70910-900 Brasília, DF, Brasil.

De acordo com [7], 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 do Sul e Centro-Oeste. Quanto ao rebanho bovino, entre 1970 e 2010, a região Centro-Oeste e Norte tinha 24% do rebanho nacional, valor 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 podem ter características semelhantes, não fazendo sentido analisá-los em separado. Já as micro e mesorregiões podem ser grandes o suficiente gerando grupos territoriais com características distintas.

Uma técnica comumente utilizada para identificar grupos homogêneos de unidades observacionais é a análise de *clusters* (ou de agrupamentos). A partir de variáveis caracterizando cada unidade, essa metodologia busca identificar subgrupos similares de observações. Com isso, é possível: (a) a construção de conjuntos homogêneos de municípios, por exemplo, que podem receber tratamento similar em termos de políticas públicas; (b) 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 (AMC's) para traçar um mapa de AMC's, contiguas e não contiguas, com características homogêneas a partir de determinadas variáveis agropecuárias. Esse estudo priorizou a utilização de AMC's, pois essa divisão territorial impõe padronização na divisão geográfica das regiões ao longo do tempo, característica não necessariamente presente nas divisões políticas municipais. Segundo [10], entre 1872 e 2000, o número de municípios brasileiros saltou de 642 para 5507, sendo que não necessariamente suas fronteiras foram respeitadas. Esse autor define as Áreas Mínimas Comparáveis (AMC's) como um agregado de municípios passíveis de comparações intertemporais de forma significante, não caracterizando uma divisão política ou administrativa. Dado que estamos utilizando AMC's como 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 parte aborda a metodologia utilizada para formação dos grupos homogêneos de AMC's (*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 parte apresenta os principais resultados oriundos da análise de *clusters* empregada. A quinta parte é reservada para as conclusões do trabalho.

2 METODOLOGIA

Os algoritmos de clusterização utilizados baseiam-se na metodologia exposta em [3], [4] e [5], onde são amplamente discutidos. 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 diferentes abordagens, além de prover diferentes configurações de agrupamentos que poderão ser utilizadas em diferentes situações de políticas públicas.

Em [5] é apresentado uma descrição sucinta dos algoritmos aglomerativos de clusterização hierárquica tradicionais e as modificações neste método a fim de incorporar a restrição de unidades geográficas contíguas. Ainda, é discutido o método *k-means* para construção de *clusters* não contíguos, que necessita, a priori, que o número de agrupamentos seja determinado.

2.1 Algoritmos aglomerativos de clusterização hierárquica

Este trabalho segue a dinâmica dos algoritmos combinatórios, os quais têm uma estrutura de formação aglomerativa de *clusters* do tipo hierárquica, conforme trabalho de [9]. De maneira geral, essa estrutura segue os seguintes passos:

1. Seja uma base de N *clusters* iniciais a serem agrupadas em grupos homogêneos (por exemplo, municípios). Em geral, cada um desses N *clusters* contém apenas uma unidade inicialmente. A cada unidade i está associado um vetor de m características $x_i = [x_{i,1}x_{i,2} \ldots x_{i,m}]$, como por exemplo, socioeconômicas.

2. Calcula-se a distância entre todos os pares formados por elementos dentre esses N *clusters* iniciais. Distância, nesse caso, pode ser qualquer métrica de dissimilaridade (ou similaridade, dependendo do algoritmo) entre o conjunto de atributos $x_i = [x_{i,1}x_{i,2} \ldots x_{i,m}]$. Entre as diversas medidas de dissimilaridade possíveis, pode-se citar a medida de Ward, que permite encontrar *clusters* de tamanhos não muito diferentes[3]. Outras medidas de dissimilaridade podem ser consultadas em [3], [4] e [6].

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 $N - 2$.

[3]Conforme exposto em [3] e [4], 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.

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 (dendograma) 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.

O passo final é então selecionar o número de *clusters* ou de grupos homogêneos por meio de medidas estatísticas, como *CCC*, *pseudo-F* e *pseudo-t*^{2} (ver [9]). De maneira geral, essas medidas estão associadas a um indicador de dissimilaridade agregada entre todos os *clusters* construídos.

2.2 Algoritmos de clusterização hierárquica espacial

No contexto deste trabalho, as unidades a serem agrupadas são AMC's, que formarão *clusters* de AMC's homogêneas, para as quais políticas de desenvolvimento regional específicas possam ser propostas. Nesse caso, espera-se que os *clusters* formados agreguem AMC's homogêneas e espacialmente vizinhas.

A fim de incorporar explicitamente a restrição de contiguidade entre as AMC's que compõem um mesmo *cluster* foram propostas algumas modificações no algoritmo de clusterização, conforme a seguir:

1. Seja C uma base inicial de N unidades geográficas. Inicialmente, cada uma dessas N observações consiste em um *cluster* isoladamente e tem um conjunto de atributos $x_i = [x_{i,1}x_{i,2}\ldots x_{i,m}]$. Para cada uma dessas N unidades, encontra-se a lista de vizinhos, de acordo com algum critério espacial. Nesse caso, foram definidos como vizinhas as AMC's 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* ([1] e [2])[4].

2. Calcula-se a distância entre todos os pares formados por elementos estritamente vizinhos na lista de N unidades, segundo medida de Ward. 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] 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 [3] e [4], 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*.

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

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 AMC's 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 $x_i = [x_{i,1}x_{i,2}\ldots x_{i,m}]$.

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 para a escolha do número de agrupamentos mais apropriado. Uma explicação a respeito dos diversos critérios de seleção do número de *clusters*, comumente utilizados em *softwares* estatísticos, é apresentado em [6].

2.3 Algoritmo de agrupamentos *k-means*

Em alternativa aos algoritmos aglomerativos hierárquicos 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. Uma explicação detalhada acerca deste método pode ser consultada em [5].

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 isso, 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*.

2.4 Medidas de homogeneidade para os *clusters* formados

Uma das comparações a serem feitas 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. A medida R^2 é então dada pela expressão:

$$R^2 = 1 - \frac{WSS}{TSS} \tag{1}$$

onde, WSS (*within sum of squares*) representa a variabilidade total, calculada pelo somatório das variabilidades quadráticas dentro de cada *cluster* e, TSS (*total sum of squares*) indica o somatório total dos quadrados.

Pode-se mostrar que o 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 da variabilidade quadrática dentro de cada agrupamento k, WSS_k. Se os valores das variáveis nas AMC's em cada *cluster* fossem exatamente os mesmos, teríamos $WSS_1 = WSS_2 = \ldots = WSS_K = 0$, 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.

3 BASE DE DADOS

Neste estudo foram utilizadas as bases de dados do ano 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 site 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. 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, tais como:

a) Características físicas: considera os aspectos do solo que podem influenciar no desenvolvimento das plantas, como armazenamento de água e aeração do solo;

b) Tipos de clima: descreve o nível de umidade conforme o número de meses com baixa umidade;

c) Fertilidade: analisa os atributos químicos do solo que podem ajudar ou restringir o desenvolvimento das plantas, por exemplo, a concentração de nutrientes benéficos e a concentração de alumínio;

d) Limitações do Solo: considera as limitações naturais de cada tipo de solo, como a disponibilidade de nutrientes e característica de relevo;

e) Seca: analisa a duração e distribuição da seca ao longo dos meses;

f) Tipos de Solo: agregado de fatores que tem por objetivo avaliar o potencial do solo para uso agrícola;

g) Temperatura: classifica as áreas de acordo com as temperaturas médias;

h) Topologia: analisa e classifica o terreno de acordo com as ondulações e declives da superfície;

i) Uso do Solo: classifica o terreno de acordo com a sua utilização.

A união desses bancos gerou um conjunto de 118 variáveis de análise, número elevado para construir um modelo e com alta 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 tais que estes contabilizassem por 99% da variabilidade das 118 variáveis originais. Ao final, selecionamos 83 componentes principais – estes foram então utilizados como variáveis para as análises de agrupamentos[5]. Essa análise é melhor detalhada em [8].

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.

4.1 Identificação do número de *clusters*

A primeira metodologia abordada foi a análise de agrupamentos aglomerativos hierárquicos não contíguos. Nos dados em análise, a medida *pseudo-t*^{2} foi usada para elucidar o número de *clusters* a ser utilizado. Ao longo da sequência aglomerativa de *clusters*, pelo método hierárquico, a medida *pseudo-t*^{2} indica junções de dois *clusters* que são relativamente diferentes entre si, ou seja, a medida *pseudo-t*^{2} aponta situações de junção forçada de dois grupos de AMC's. 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 implica que podemos selecionar números diferentes de *clusters*, mesmo usando apenas um único critério (no caso, o *pseudo-t*^{2}). 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-t*^{2} 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

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

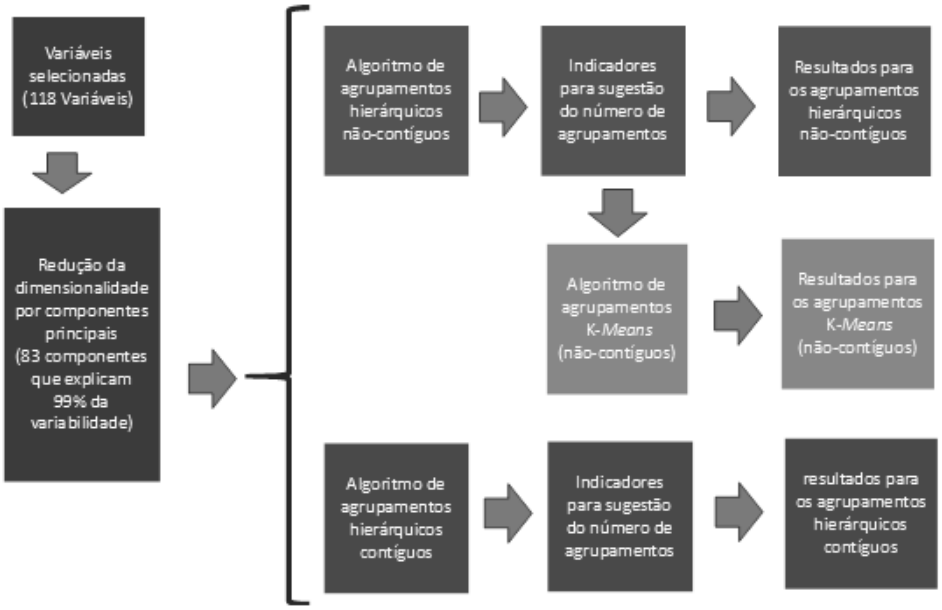

Figura 1: Análise de agrupamentos com diferentes metodologias.

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 dos *clusters* espaciais implicarem em 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.

Abaixo as Figuras 2 e 3 exibem, respectivamente, os gráficos da 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.

A partir de dez agrupamentos, os *clusters* espaciais 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 AMC's e, portanto, tende a diminuir o R^2.

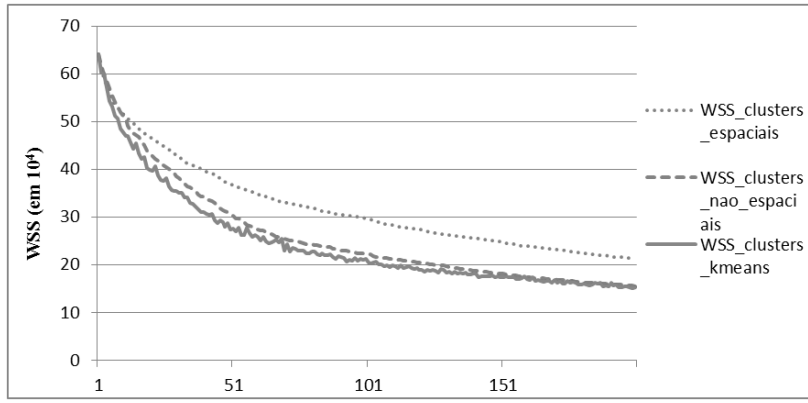

Figura 2: Comparação do *WSS* dos *clusters* para diferentes métodos.

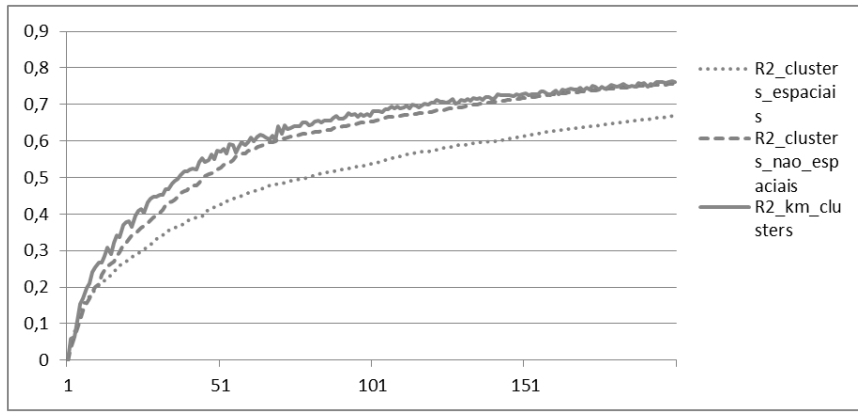

Figura 3: Comparação do R^2 dos *clusters* para diferentes métodos.

Observamos, por exemplo, que o R^2 para o método *k-means*, com vinte e um *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 do 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 da curva R^2 não ser estritamente crescente para o caso *k-means*, que ocorre devido ao fato de tal método não ser sequencial, ao contrário dos dois métodos hierárquicos.

Dentre os quatro valores para o número de *clusters*, sugeridos acima, decidiu-se utilizar o valor igual a vinte e um *clusters* para os casos não espaciais. 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 vinte e um *clusters* para o método *k-means*, aproximadamente 38%, permitindo assim certo grau de comparação entre os métodos.

4.2 Resultados para *clusters* por métodos não espaciais

Como mencionado anteriormente, tanto para o método aglomerativo hierárquico não espacial quanto para o método *k-means* foram considerados o mesmo número de agrupamentos. Entretanto, nas análises dos resultados, o foco foi mais concentrado no método *k-means*, devido ao fato dos agrupamentos deste método terem apresentado um maior R^2. Estas análises são feitas a partir dos *box-plots* gerados, 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 por utilizar a mediana, como medida de centralidade, por ela ser robusta a observações extremas dentro de cada *cluster*.

Dentre 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 que na agricultura são relevantes as culturas de sorgo, algodão, soja, milho, feijão e arroz.

A Figura 4 exibe a posição do *cluster* 5 no território brasileiro. Nota-se que esse *cluster* é formado por AMC's presentes nos estados do Mato Grosso (MT), Mato Grosso do Sul (MS), Goiás (GO), Minas Gerais (MG) e Bahia (BA).

Outro *cluster* que se destaca é o *cluster* 1 (Figura 5), localizado primordialmente, no estado do Pará (PA). Primeiramente, apesar de não se tratar de um método espacial, as unidades que o compõem são bastante próximas, apresentando contiguidade entre todas as AMC's, menos uma. Na produção agrícola, as culturas de maior destaque foram arroz e mandioca, enquanto que, na pecuária, as principais criações foram as de bovinos e equinos.

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*.

Figura 4: *Cluster* número 5 – método k-means.

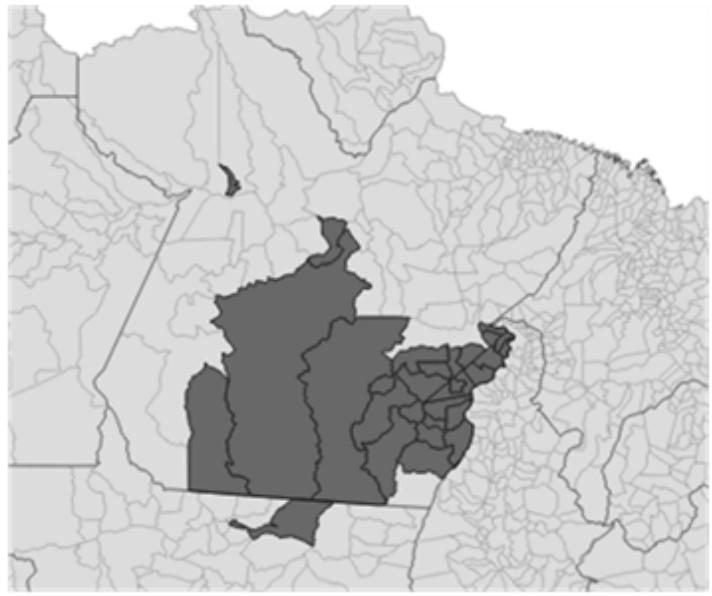

Figura 5: *Cluster* número 1 – método k-means.

4.3 Resultados para *clusters* por métodos espaciais

Para o caso de clusterização espacial, considera-se no modelo a contiguidade entre as AMC's. 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 AMC's pode ocorrer a formação de *clusters* pequenos. Com o algoritmo hierárquico espacial, metodologia encontrada em [3] e [4], os *clusters* contíguos são definidos para utilização nessa análise. Dentre os números de *clusters* sugeridos pelos indicadores, optou-se por utilizar quarenta agrupamentos (Figura 6), já que o R^2 nesse modelo é bastante próximo do R^2 para o método *k-means* com vinte e um *clusters*, discutido na subseção anterior.

Figura 6: Clusters – método espacial.

Considerando os *box-plots* para o caso de clusterização espacial, realizou-se uma análise gráfica das principais culturas e criações levadas em consideração no estudo. Desta forma, pode-se observar por meio da mediana (medida de centralidade) os principais valores não discrepantes das variáveis dentro de cada *cluster* analisado.

Dentre 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; iv) arroz, com mediana maior; 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.

A Figura 7 exibe a disposição do *cluster* no Brasil, sendo ele localizado totalmente no estado do Mato Grosso (MT). Por ser um agrupamento que demonstrou alta relevância nas culturas e criações, destacam-se as AMC's 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.

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 (vizinhos do *cluster* 4) 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, pode estimular o desenvolvimento de *clusters* próximos que se beneficiem da produção dessa matéria-prima. A Figura 8 exibe o posicionamento destes *clusters* e a relação de proximidade entre eles.

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*.

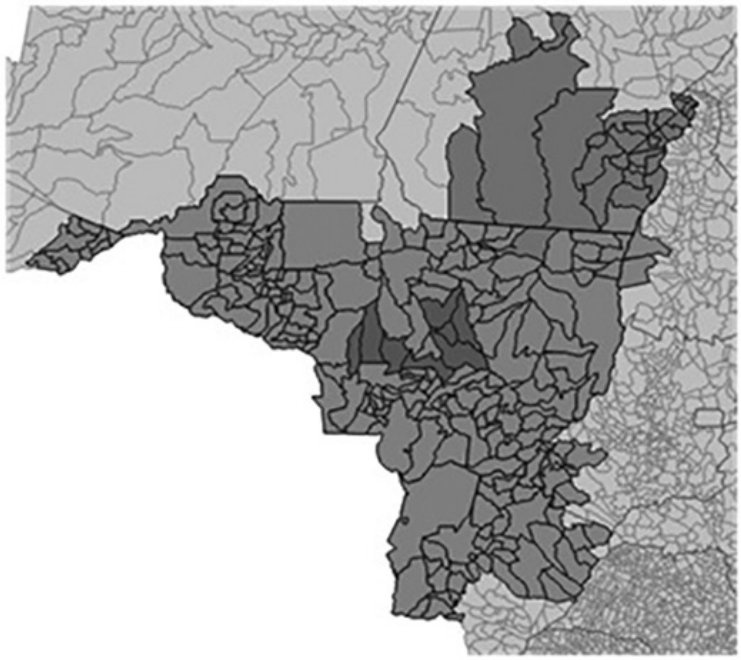

Figura 8: *Clusters* 3, 4 e 6 – método espacial.

5 CONSIDERAÇÕES FINAIS

Neste artigo, o objetivo foi a identificação de agrupamentos de Áreas Mínimas Comparáveis (AMC's) para a análise de variáveis de produção agropecuária, variáveis físicas e variáveis climáticas, permitindo se traçar 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, situação quando uma área de estudo é suficientemente grande e com características muito genéricas, de forma que não reflita a realidade observada, ou 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-*t^2, 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 dos métodos *k-means* e hierárquico espacial.

Para o método *k-means*, além de uma visão geral de sua aplicação, ressaltou-se o grupo de AMC's incluindo Unaí (MG), Nova Mutum (MT), Sorriso (MT), Rio Verde (GO), e mais vinte

e sete AMC's, que compõem o *cluster* 5, pelo fato deste apresentar maior mediana em algumas das culturas consideradas, em especial, soja, milho, feijão, arroz, algodão e sorgo. Para a clusterização espacial, o grupo de AMC's incluindo Campo Novo do Parecis (MT), Campos de Júlio (MT), Diamantino (MT), Lucas do Rio Verde (MT), Nova Mutum (MT), Sapezal (MT), Sorriso (MT) e Tapurah (MT) recebeu mais foco. Ao observar os mapas contendo os *clusters*, nos deparamos com o fato das AMC's do agrupamento incluindo Nova Mutum (MT), Sapezal (MT), Sorriso (MT) e Tapurah (MT) para o método espacial estarem contidas no agrupamento incluindo Unaí (MG), Nova Mutum (MT), Sorriso (MT) e Rio Verde (GO) para o método *k-means*. Por um lado, este ponto reafirma o conceito de clusterização e, por outro, ele 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 AMC's brasileiras, de forma a servir como guia para políticas públicas de desenvolvimento da agropecuária, podendo levar a políticas públicas direcionadas e espacializadas. Ainda, utilizando-se outros conjuntos de variáveis, as metodologias discutidas aqui poderão ser aplicadas a outros temas de interesse para análise.

ABSTRACT. This paper presents a clustering analysis of Minimum Comparable Areas (MCAs) to draw a map of homogeneous grouping from a combination of climatic variables, soil characteristics and agricultural production. The methodology allows the visualization of interactions among the many different variables used, indentifying, for example, coexistence patterns, at the municipal level, of different crops. The discussion presents the traditional algorithms with no contiguity (hierarchical algorithm and k-means) and the agglomerative hierarchical algorithm with contiguity. Therefore, this paper seeks to explore differences among the typologies built with different approaches, as well as, provide alternative configurations of grouping. Also, the methodologies discussed allow the incorporation of traditional criteria for choosing the number of clusters, such as the CCC, pseudo-F and pseudo-t^2 statistics.

Keywords: Spatial clustering, hierarchical algorithms, k-means.

REFERÊNCIAS

[1] L. Anselin. "Spatial econometrics: methods and models", Dordrecth: Kluwer Academic, (1988).

[2] L. Anselin & R. Florax. "Advances in spatial econometrics", Heidelberg: Springer-Verlag (2000).

[3] A.X.Y. Carvalho et al. Spatial hierarchical clustering, *Revista Brasileira de Biometria*, **27**(3) (2009), 412–443, São Paulo.

[4] A.X.Y. Carvalho et al. Clusterização hierárquica espacial com atributos binários, *Revista Brasileira de Biometria*, **19**(1) (2011), 147–197, São Paulo.

[5] A.X.Y. Carvalho et al. Clusterização espacial e não espacial: um estudo aplicado à agropecuária brasileira, *Coleção Texto para Discussão n*°2279*, Ipea*, Brasília (2017).

[6] M. Charrad et al. Nbclust: an *r* package for determinig the relevant number of clusters in a data set. *Journal of Statistical Software*, **61**(6) (2014), 1–36.

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

[8] T. Hastie, R. Tibshirani & J. Friedman. "The elements of statistical learning: data mining, inference and prediction", Standford: Springer (2001).

[9] R. Khattree & D. N. Naik. "Multivariate data reduction and discrimination with SAS software", Cary: SAS Institut, (2000).

[10] E. Reis, M. Pimentel & A.I. Alvarenga. Áreas mínimas comparáveis para os períodos intercensitários de 1872 a 2000, *Ipea/Dimac*, (2008), Mimeografado, Rio de Janeiro.