ABORDAGEM HÍBRIDA MULTICRITÉRIO – MINERAÇÃO DE DADOS APLICADA A CLASSIFICAÇÃO DE UNIDADES DA FEDERAÇÃO COM BASE NA POPULAÇÃO ECONOMICAMENTE OCUPADA

Glauco Barbosa da Silva
Centro de Análises de Sistemas Navais (CASNAV)
Praça Barão de Ladário, s/n, Centro, Rio de Janeiro, Brasil
glaucos@id.uff.br

Marta Duarte de Barros
Universidade Federal Fluminense
Rua Passos da Pátria, 156, Bloco D, sala 309, Niterói, Brasil
marta_uff@hotmail.com

Helder Gomes Costa
Universidade Federal Fluminense
Rua Passos da Pátria, 156, Bloco D, sala 309, Niterói, Brasil
helder.hgc@gmail.com

Resumo

Por meio de uma combinação de técnicas de mineração de dados e métodos multicritério, o presente trabalho tem como objetivo propor um novo agrupamento de Unidades da Federação (UFs) por Atividade Principal da População Ocupada, não observando a conectividade geográfica das UFs. A partir de uma amostra baseada nos dados do censo demográfico brasileiro(2010), assumindo cada unidade da federação como uma instância e cada atividade da população ocupada como um atributo, excluindo-se as classes originais (regiões), a base de dados é submetida a um método clusterização, que gera as classes iniciais para o método de composição probabilística de preferências (CPP-TRI). Com base nas similaridades e dissimilaridades das atividades econômicas da população ocupada, evidenciadas pelo processo de formação de clusters com a ferramenta WEKA e na classificação feita pelo método CPP-TRI, novos agrupamentos são identificados e analisados. A análise exploratória dos novos grupamentos permitiu ampliar o conhecimento acerca das especificidades das UFs com base quantitativa e viés econômico. A identificação de novos blocos econômicos, ocultos quando da utilização dos modelos tradicionais fundamentados em critérios geográficos, pode possibilitar a formação e o desenvolvimento de ações de cooperação entre as UFs.

Palavras-Chaves: CPP-TRI; MCDA; *k-means*.

1. INTRODUÇÃO

Com dimensões continentais, o Brasil é um país que além das distâncias físicas ao longo do território, possui grandes distâncias sociais, culturais e econômicas. Com a missão de investigar as características da população brasileira, o Instituto Brasileiro de Geografia e Estatística (IBGE) realiza o censo demográfico, que constitui importante fonte para ampliar o conhecimento acerca das especificidades do país.

Em geral, os resultados obtidos por meio do censo estão agrupados por macrorregiões, que seguem uma divisão política e geográfica, abordagem que pode mascarar aspectos importantes e diferenças intrarregiões. Além disso, considerando os avanços alcançados na última década, é de se esperar que a ocupação da população tenha sofrido um aumento significativo e esteja ainda mais diversificada.

A partir dos dados disponíveis do censo demográfico (2010) é nebulosa a ideia de como as Unidades da Federação (UF) poderiam ser agrupadas, ou seja, a pergunta básica que se apresenta é: tomando-se por base as similaridades e dissimilaridades das atividades econômicas da população ocupada, que blocos de UFs podem ser formados?

Durante as últimas décadas, os avanços das tecnologias em bases de dados aliados à redução do custo de armazenamento têm possibilitado o aumento da coleta e armazenagem de grandes volumes de dados, oriundos de diferentes áreas de aplicação, tais como: serviços bancários, telefonia, processamento de imagens, meteorologia, médicos, entre outros. Quando bem processados, estes dados representam um valioso conhecimento. Entretanto, o grande volume de dados ultrapassou a capacidade humana de analisá-los em tempo oportuno, ocasionando um problema comum: muitos dados, mas pouco conhecimento sobre eles. Dessa forma, surge a necessidade do uso de ferramentas capazes de auxiliar na análise dos dados.

Considerando a complexidade e o volume dos dados coletados pelo censo demográfico, a mineração de dados apresenta-se como uma ferramenta candidata a revelar aspectos relevantes não conhecidos até então e potencialmente úteis.

Dentre os diferentes tipos de tarefas que podem ser executadas no processo de mineração de dados, tem-se: as regras de associação, hierarquias de classificação, clusterização, padrões seqüenciais e os padrões em séries temporais. Witten, Frank e Hall [1] descrevem detalhadamente diferentes técnicas e tarefas.

Uma das técnicas de mineração de dados que proporcionam um melhor entendimento da base de dados é a clusterização, que está normalmente associada com a análise exploratória, pois envolve problemas em que há pouca informação (ou pouco se conhece) a priori acerca dos dados.

Sant'Anna, Costa e Pereira [2] propõe um método multicritério de classificação ordenada baseado em composição probabilística que prescinde completamente da atribuição de pesos aos critérios - CPP-TRI - que permite levar em conta a imprecisão nas avaliações para produzir medidas probabilísticas de distância de cada alternativa a ser classificada a cada classe.

Assumindo que no Brasil pouco se conhece da relação entre as UFs, no que tange a atividade econômica principal da população ocupada, por meio de uma combinação de técnicas de mineração de dados com auxilio da ferramenta Weka e o método multicritério CPP-TRI, o presente trabalho tem como objetivo propor um novo agrupamento de UFs, não observando a conectividade geográfica entre elas.

Waikato Environment for Knowledge Analysis (Weka) é uma ferramenta desenvolvida pela Universidade de Waikato na Nova Zelândia, que reúne algoritmos de

aprendizagem de máquina e técnicas de pré-processamento de dados Witten, Frank e Hall [1]. Além dos recursos implementados e da facilidade na utilização, a ferramenta Weka foi escolhida por ser um software livre.

Como trabalhos correlatos, Kubrusly [3] faz uma análise da população ocupada focando nas diferenças das atividades econômicas e regiões, numa abordagem com víeis na renda da população. Sant'Anna, Faria e Costa [4] empregam a composição probabilística de preferências para classificar em categorias ordenadas, e o método das k-médias para classificar os municípios fluminenses quanto ao cumprimento da disposição constitucional de disponibilidade de creches para a população infantil.

Este trabalho está estruturado em cinco seções distribuídas da seguinte forma: na introdução são apresentados a descrição do problema da pesquisa, a questão de pesquisa e o objetivo e a estrutura do trabalho; a seção 2 insere a fundamentação teórica com conceitos e definições sobre Mineração de Dados e multicritério com a finalidade de possibilitar um melhor entendimento da elaboração do instrumento metodológico empregado; na seção 3 são apresentados os dados utilizados no estudo; na seção 4 é descrita a metodologia empregada, destacando os critérios para a seleção dos dados, as etapas do processo, a construção dos experimentos e a análise dos resultados. Encerra-se com a seção 5, que apresenta a conclusão alinhada com os objetivos da pesquisa e sugestões para trabalhos futuros.

2. CONCEITOS E DEFINIÇÕES

Nesta seção, são descritos os fundamentos teóricos e conceituais, que serviram de base para a pesquisa.

2.1. MINERAÇÃO DE DADOS E K-MEANS

Segundo Witten, Frank e Hall [1], a Mineração de Dados é uma das etapas do processo de descoberta do conhecimento em banco de dados (KDD - Knowledge Discovery from Data). O KDD consiste de um processo essencial, onde técnicas estatísticas e computacionais são aplicadas para identificar padrões em grandes volumes de dados.

Segundo Steiner *et al.* [5], o processo de KDD é um conjunto de atividades contínuas, que compartilham conhecimento descoberto a partir de bases de dados. É um processo não trivial de identificação de padrões válidos, novos, potencialmente úteis e compreensíveis em dados, composto das seguintes etapas: Seleção, Pré-processamento e Limpeza, Transformação, Mineração de Dados, Interpretação e Análise Exploratória dos dados (Fayyad, Piatetsky-Shapiro e Smyth [6]).

As tarefas de mineração de dados podem ser classificadas em duas categorias: descritivas – quando se deseja descrever e/ou entender características importantes dos dados, padrões e regras, com os quais se está trabalhando (sumarização, análise de associações, clusterização); e preditivas – quando se deseja prever o valor desconhecido de um determinado atributo, a partir da análise histórica dos dados armazenados na base (previsões e classificação), com diferentes aplicações.

A clusterização é uma tarefa bastante utilizada na mineração de dados, que particiona um conjunto de dados conforme a similaridade desses dados, isto é, os grupos formados buscam por grande similaridade intragrupo (homogeneidade) e grande dissimilaridade entregrupos (heterogeneidade).

Segundo Witten, Frank e Hall [1], a clusterização é utilizada quando as instâncias não apresentam classes previamente definidas - base não-supervisionada. Para medir a similaridade ou a dissimilaridade entre os dados, normalmente, utiliza-se a como métrica a

distância euclidiana entre objetos de dados, quando se está trabalhando com variáveis intervalares.

Os principais métodos de clusterização podem ser agrupados em duas categorias: Hierárquicos e Não-Hierárquicos. Dentre os métodos não-hieráquicos, o mais popular e utilizado é método *k-means*, que foi selecionado para a clusterização das UFs por atividade econômica da população ocupada.

O *k-means* é um método que busca particionar um conjunto de dados(D) com *n* componentes em k grupos (clusters – $C_1,...,C_k$), onde : $k \leq n$; $C_i \subset D$ ^ $C_i \cap C_j = \emptyset$; para ($1 \leq i, j \leq k$). Seja $\underline{\boldsymbol{c_i}}$ o centróide do cluster C_i e $\boldsymbol{p}$ um elemento de D, tal que $\boldsymbol{p} \in C_i \rightarrow d_i$ é a menor distancia entre $\boldsymbol{p}$ e qualquer dos centróides. O centróide do cluster é representado pela média dos elementos do cluster. A qualidade dos clusters pode ser medida pela soma do erro quadrático (E) entre os p elementos e o centróide $\underline{\boldsymbol{c_i}}$ (distancia Euclidiana – $d(p, c_i)$), ou seja, a variação intracluster. Segundo Han, Kamber e Pei [7], em termos gerais, o algoritmo pode ser descrito como:

Algoritmo k-means

Input:

k: número de clusters.

*D: Conjunto de dados contendo **n** objetos.*

Output:

Um conjunto de k clusters.

Método:

(1) Escolhe arbitrariamente k objetos de D como centróides iniciais;

(2) Repetir

(3) Alocar cada elemento de D ao cluster cujo centróide esteja mais próximo (mais similar – distância Euclidiana);

(4) Recalcular a média dos objetos alocados no cluster e determinar o novo centróide para cada cluster;

(5) até que nenhum elemento mude de cluster;

2.2. MÉTODO DE CLASSIFICAÇÃO ORDENADA BASEADO EM COMPOSIÇÃO PROBABILÍSTICA (CPP-TRI)

Baseado na transformação probabilística proposta por Sant'Anna [8], e conforme Sant'Anna, Costa e Pereira [2], as avaliações numéricas iniciais segundo cada critério são tratadas como parâmetros de locação de distribuições de probabilidades. Segue princípios da modelagem econométrica clássica, assumindo, além de distribuições normais (cujas variâncias são estimadas a partir dos dados disponíveis), idêntica distribuição e independência entre as perturbações que provocam a imprecisão nas medidas. Ou seja, a avaliação segundo cada critério é representada por uma distribuição de probabilidade, possibilitando o calculo da probabilidade de cada alternativa ter uma avaliação acima ou abaixo do perfil representativo de cada classe, segundo cada critério.

Das probabilidades de sobreclassificação, segundo cada critério, são derivadas as classificações globais sem a necessidade de atribuição de pesos.

Em termos gerais, maiores detalhes podem ser obtidos em Sant'Anna, Costa e Pereira [2], o método pode ser descrito como a seguir:

- G é uma família formada por m critérios;

- A é um vetor do R^m que guarda as avaliações de uma alternativa A, segundo cada um dos m critérios em G.
- C é um conjunto de r classes ordenadas, nas quais as alternativas serão classificadas;
- Estimação da Variância

A variância deve, em princípio, ser suficientemente grande para que a probabilidade da alternativa pertencer mesmo às classes com valores segundo o k-ésimo critério mais afastados de a_k não seja tão pequena que torne inútil a informação dada pelos outros critérios, ou seja, nenhum critério tenha sozinho o poder de veto a nenhuma classificação. Este princípio pode ser satisfeito, na prática, estimando a variância das medidas segundo cada critério pela variância do conjunto de medições registradas segundo este critério para várias alternativas a classificar ou para os vários perfis oferecidos como representativos das classes.

- Cálculo das Probabilidades de Preferência segundo cada Critério Isoladamente

Sejam A_{ik+} e A_{ik-} as probabilidades da alternativa A apresentar valor respectivamente acima e abaixo dos valores informados para o critério k-ésimo nos perfis da classe i-ésima. Por independência entre as perturbações afetando diferentes alternativas e diferentes perfis,

$$A_{ik+} = \Pi_j P[X_k > Yijk] \quad \text{(eq.1)}$$

$$A_{ik-} = \Pi_j P[Xk < Yijk] \quad \text{(eq.2)}$$

- Avaliação da Credibilidade das Relações

Supondo independência entre as avaliações por diferentes critérios, as probabilidades conjuntas de estar acima e abaixo de todos os perfis da classe por todos os critérios são dadas por

$$A_{i+} = \Pi_k A_{ik+} \quad \text{(eq.3)}$$

$$A_{i-} = \Pi_k A_{ik-} \quad \text{(eq.4)}$$

- Classificação

A classificação é baseada na comparação das diferenças ($A_{i+} - A_{i-}$). Definidos os perfis de modo que as classes estão ordenadas em ordem crescente, a regra de classificação consiste em: alternativa A pertence à classe C_i para a qual essa diferença é mais próxima de zero. Identifica-se o menor valor absoluto para a diferença ($A_{i+} - A_{i-}$), a alternativa é alocada à classe i; em caso de empate, a alternativa pode ser classificada em duas classes adjacentes.

3. DESCRIÇÃO DOS DADOS

Por meio do censo demográfico, o Instituto Brasileiro de Geografia e Estatística (IBGE) investiga as características da população e dos domicílios do território brasileiro, constituindo uma relevante fonte de referência para o conhecimento das condições de vida da

população em todos os municípios e em seus recortes territoriais internos.

Em geral, os dados do censo são apresentados no documento Resultados Gerais da Amostra (RGA), que dispõe de tabelas de resultados, notas técnicas e uma análise dos aspectos divulgados. Os resultados estão organizados em tabelas para Brasil, Grandes Regiões, Unidades da Federação e municípios, compreendendo informações sobre aspectos das pessoas com deficiência; migração; nupcialidade; fecundidade e mortalidade infantil; educação; trabalho e rendimento; deslocamento para trabalho e estudo; e domicílios. A seção trabalho abrange, entre outros, informações da taxa de atividade (percentual de pessoas economicamente ativas na população), o nível da ocupação (percentual de pessoas ocupadas na semana de referência) e a distribuição de pessoas ocupadas por atividade econômica.

O detalhamento das técnicas de coletas de dados estatísticos não é a ênfase deste trabalho e pode ser consultado em IBGE [9]. Uma análise apresentada no RGA mostra que em quatro dos vinte e um agrupamentos das seções de atividade do trabalho principal, inseriam-se praticamente a metade (50,3%) da população economicamente ocupada, são eles: **APPFPA** - Agricultura, Pecuária, Produção Florestal, Pesca e Aqüicultura; **ITR** - Indústria de Transformação; **CONST** – Construção; **COM** – Comércio, Reparação de Veículos Automotores e Motocicletas).

Além dessas atividades, para o presente estudo, foram consideradas as atividades que apresentaram percentuais, a nível nacional, superiores a 1%, são elas: **ADMP** – Administração pública, defesa e seguridade social; **EDU** – Educação; **TRANS** - Transporte, armazenagem e correio; **ALIM** - Alojamento e alimentação; **INF** - Informação e comunicação; **FINAN** - Atividades financeiras, de seguros e serviços relacionados; **CIEN** - Atividades profissionais, científicas e técnicas; **ADMC** - Atividades administrativas e serviços complementares; **SAU** - Saúde humana e serviços sociais; **OUT** - Outras atividades de serviços; **SVCD** – Serviços Domésticos; totalizando 15 atributos de interesse e 27 instâncias(UF). Conforme apresentado na figura 1, na distribuição da população brasileira de 10 anos ou mais de idade, ocupadas por atividade principal, a atividade **COM** apresenta o maior o percentual com 18,62%; **INF** apresenta o menor percentual com 1,41%.

A tabela I consolida o percentual das pessoas ocupadas por atividades econômicas nas grandes regiões por ocasião do censo; por exemplo, para o total de pessoas ocupadas na atividade APPFPA encontram-se: 11,77% na região Norte; 41,12% na região Nordeste; 22,97% na região Sudeste; 17,61% na região Sul; e 6,53% na região Centro-Oeste. A partir dos dados consolidados é possível identificar quais são as regiões com maior percentual de pessoas ocupadas para cada atividade. Entretanto, é importante destacar que os percentuais regionais não são uniformemente distribuídos. Para a atividade APPFPA na região Sudeste (22,97%), por exemplo, tem-se Minas Gerais (MG) com 12,24%, Espírito Santo (ES) com 2,44%, Rio de Janeiro (RJ) com 1,25% e São Paulo (SP) com 7,04%; que corrobora a hipótese da diversidade intrarregional.

Como síntese da análise presente no RGA, pode-se destacar que esta procura identificar as marcantes distinções regionais, refletida na distribuição da população ocupada. Contudo, o mesmo não destaca as similaridades das atividades desenvolvidas pelas diferentes Unidades da Federação (UF).

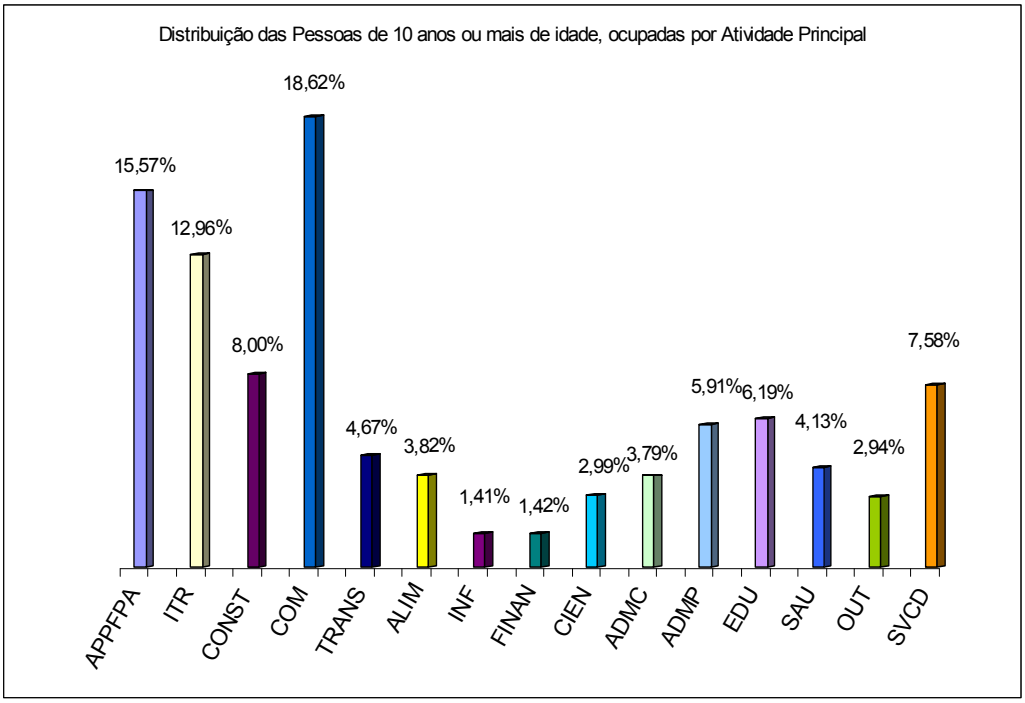

Fig. 1. Distribuição da População Brasileira Ocupada por Atividade

TABLE I
PERCENTUAL DA POPULAÇÃO OCUPADA POR GRANDES REGIÕES

		Norte	Nordeste	Sudeste	Sul	Centro-Oeste
Seção de atividade do trabalho principal	**APPFPA**	11,77%	41,12%	22,97%	17,61%	6,53%
	ITR	4,38%	16,69%	49,73%	22,90%	6,31%
	CONST	7,06%	23,83%	44,70%	15,91%	8,49%
	COM	7,30%	23,68%	44,25%	16,36%	8,40%
	TRANS	6,56%	20,15%	49,88%	16,33%	7,08%
	ALIM	6,56%	22,29%	48,12%	14,66%	8,37%
	INF	3,44%	13,12%	58,82%	16,35%	8,26%
	FINAN	3,39%	14,51%	57,19%	16,78%	8,12%
	CIEN	4,58%	14,35%	55,64%	17,47%	7,96%
	ADMC	5,57%	17,87%	54,69%	14,07%	7,79%
	ADMP	9,50%	25,67%	39,35%	13,85%	11,63%
	EDU	7,62%	26,81%	42,86%	14,86%	7,85%
	SAL	5,71%	21,13%	50,62%	15,28%	7,26%
	OUT	6,23%	21,00%	48,45%	15,47%	8,84%
	SVCD	6,70%	23,36%	47,76%	13,36%	8,82%

Fonte: IBGE, 2010. (Adaptado pelo autor)

4. METODOLOGIA

4.1. SELEÇÃO E PREPARAÇÃO DOS DADOS

Dentre as 23 atividades presentes no RGA foram excluídas as atividades classificadas como maldefinidas (aproximadamente 6%) e as atividades que representavam um percentual menor que 1% em termos de Brasil, foram elas: Indústrias extrativas (0,46%); Eletricidade e Gás (0,25%); Água, Esgoto, Atividades de Gestão de Resíduos e Descontaminação (0,67%); Atividades Imobiliárias (0,43%); Artes, Cultura, Esporte e Recreação (0,83%); e Organismos Internacionais e Outras Instituições Extraterritoriais (0,005%).

Para as 15 atividades selecionadas, foram recalculados os percentuais. Numa abordagem com foco na representatividade de cada atividade para cada UF, as distribuições das ocupações da população em cada UF foram calculadas. Por exemplo, para a instância 1 (Acre), 22,41% da população da UF está ocupada na atividade APPFPA; 4,73% na atividade ITR; 8,24% em CONST; 15,96% em COM; 3,59% em TRANS; 3,17% em ALIM; 0,74% em INF; 0,73% em FINAN; 2,08% em CIEN; 3,57% em ADMC; 10,51% em ADMP; 8,35% em EDU; 3,79% em SAU; 2,61% em OUT; e 9,52% em SVCD.

Os resultados para a seleção de atividades obtidos foram convertidos para o formato de entrada na ferramenta Weka(.arrf) e serão utilizados nas etapas seguintes.

4.2. MINERAÇÃO DE DADOS – CLUSTERIZAÇÃO

Tomando por base as 15 atividades selecionadas e as 27 unidades da federação, de modo a possibilitar a clusterização, as classes (macrorregiões) foram excluídas para a aplicação do *algoritmo k-means*. Como parâmetros de entrada (*inputs*) foram definidos: k=5, mesmo número de macrorregiões, distância euclidiana e demais parâmetros *default* da ferramenta Weka. Como resultado, o cluster 0 recebeu uma unidade; o cluster 1 recebeu 6 unidades; o cluster 2 recebeu 3 unidades; o cluster três recebeu 8 unidades; e o cluster 4 recebeu 9 unidades. O quadro 1 apresenta os novos clusters formados e tabela II apresenta as características dos centróides formados pela etapa da clusterização.

QUADRO 1. RESULTADO DO ALGORITMO K-MEANS(K=5)

Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Acre	Amapá	Distrito Federal	Espírito Santo	Alagoas
	Maranhão	Rio de Janeiro	Goiás	Amazonas
	Paraíba	São Paulo	Mato grosso	Bahia
	Piauí		Mato grosso do Sul	Ceará
	Roraima		Minas Gerais	Pará
	Tocantins		Paraná	Pernambuco
			Rio Grande do Norte	Rio Grande do Sul
			Rondônia	Santa Catarina
				Sergipe

TABELA II. CARACTERÍSTICAS DOS CLUSTERS

			Clusters - Centroids e Std-Dev									
	Full Data		C0		C1		C2		C3		C4	
	M	σ	μ	Σ	μ	σ	μ	Σ	μ	σ	μ	σ
APPFPA	19,63%	8,04%	22,41%	0	24,06%	7,94%	3,15%	1,47%	18,24%	3,99%	23,12%	4,82%
ITR	9,84%	4,55%	4,73%	0	5,73%	1,49%	10,84%	6,65%	11,07%	2,71%	11,73%	5,15%
CONST	8,06%	0,80%	8,24%	0	8,43%	0,69%	7,98%	1,05%	8,59%	0,52%	7,34%	0,55%
COM	18,43%	1,31%	15,96%	0	18,03%	1,71%	19,17%	1,09%	19,18%	0,94%	18,04%	0,91%
TRANS	4,15%	0,77%	3,59%	0	3,30%	0,60%	5,35%	0,90%	4,29%	0,37%	4,25%	0,48%
ALIM	3,60%	0,58%	3,17%	0	3,15%	0,49%	4,62%	0,35%	3,76%	0,49%	3,49%	0,27%
INF	1,06%	0,59%	0,74%	0	0,66%	0,14%	2,51%	0,45%	1,06%	0,21%	0,88%	0,31%
FINAN	1,10%	0,61%	0,73%	0	0,69%	0,05%	2,53%	0,81%	1,10%	0,22%	0,94%	0,27%
CIEN	2,41%	0,86%	2,08%	0	1,71%	0,29%	4,23%	0,10%	2,70%	0,41%	2,06%	0,60%
ADMC	3,21%	1,09%	3,57%	0	2,24%	0,66%	5,69%	0,34%	3,15%	0,32%	3,05%	0,62%
ADMP	7,64%	3,14%	10,51%	0	9,74%	4,12%	9,41%	6,39%	6,61%	1,16%	6,27%	1,20%
EDU	6,82%	1,36%	8,35%	0	8,66%	1,44%	6,66%	0,77%	6,13%	0,66%	6,09%	0,57%
SAU	3,85%	0,68%	3,79%	0	3,91%	0,58%	5,28%	0,48%	3,71%	0,43%	3,49%	0,39%
OUT	2,76%	0,42%	2,61%	0	2,47%	0,28%	3,62%	0,35%	2,90%	0,29%	2,56%	0,11%
SVCD	7,41%	1,09%	9,52%	0	7,22%	0,61%	8,96%	0,91%	7,53%	0,81%	6,68%	0,87%

Os resultados da clusterização, características dos clusters, servirão como perfis de entrada para emprego do método CPP-TRI.

4.3. APLICAÇÃO DO CPP-TRI

Para a aplicação do CPP-TRI foram considerados como perfis de referência os vetores representados pelos centróides dos clusters gerados pelo *k-means* e apresentados na tabela II.

Para que a condição de classes ordenadas fosse atendida, os valores para a atividade APPFA foram multiplicados por (-1); e foram excluídas as atividades ITR, CONST, COM, ADMP, EDU, SAU e SVCD. O Cluster 0 é composto apenas pelo Acre, sendo assim, não foi considerada na aplicação do método. Após essas transformações, os clusters são ordenados como: $C_2 > C_3 > C_4 > C_1$.

Com o auxilio de uma planilha Excel, para cada cluster foram calculadas as diferenças entre A_{i+} e A_{i-}, os resultados obtidos encontram-se consolidados na tabela III; com base nessas diferenças, as unidades foram classificadas como apresentado no quadro 2.

Comparando-se a classificação gerada pelo CPP-TRI e os clusters formados pelo *k-means*, 65% das unidades foram classificadas coincidentemente dentre dos clusters; 30% foram classificadas em cluster adjacentes.

TABELA III. DIFERENÇAS ENTRE AI+ /AI- PELO CPP-TRI.

	C1	C2	C3	C4
Amapá	-0,109054194	0,846028027	0,001348848	-0,005983642
Amazonas	-0,01267669	0,649723096	5,03724E-05	-0,00010198
Pará	-0,00042062	0,884370857	0,417823207	0,098800612
Rondônia	-0,045935043	0,900044298	0,131980116	0,031011614
Roraima	-0,000138951	0,988140337	0,404513463	0,045490127
Tocantins	-0,006990988	0,966686229	0,398241119	0,002030061
Alagoas	-0,004816571	0,959858316	0,653567828	0,163232467
Bahia	-0,03623022	0,908177449	0,325824281	0,014834241
Ceará	-0,086977336	0,954430984	0,287911626	0,013576693
Maranhão	0,085220468	0,971344909	0,957795527	0,610753848
Paraíba	0,004336674	0,976003472	0,83223238	0,392782864
Pernambuco	-0,239364994	0,793473619	0,001200201	-0,019266311
Piauí	0,001681338	0,984745343	0,926887334	0,56875098
Rio Grande do Norte	-0,414962835	0,436211871	0,003428064	-0,012600721
Sergipe	-0,025017525	0,790992166	0,069807599	0,003234167
Minas Gerais	-0,456785988	0,765664017	0,00465235	-0,061578855
Espírito Santo	-0,507603228	0,704750952	-0,010208241	-0,234650638
Rio de Janeiro	-0,996727926	-0,003647851	-0,988814704	-0,999669177
São Paulo	-0,980046786	0,056525596	-0,719270572	-0,995213898
Paraná	-0,546738521	0,664204006	-0,031604695	-0,227304777
Santa Catarina	-0,445114533	0,810269545	-0,005341898	-0,112729297
Rio Grande do Sul	-0,233755855	0,74285558	-0,001508406	-0,017427322
Mato Grosso do Sul	-0,371405707	0,864725768	0,007526527	-0,009967792
Mato Grosso	-0,425049659	0,80374645	0,026437479	-0,047844367
Goiás	-0,83963378	0,561673318	-0,034698804	-0,166104529
Distrito Federal	-0,958262308	-0,002649208	-0,540604261	-0,600247216

QUADRO 2. CLASSIFICAÇÃO RESULTANTE DO CPP-TRI.

C2	C3	C4	C1
Rio de Janeiro	Amapá	Rondônia	Pará
São Paulo	Amazonas	Tocantins	Roraima
Distrito Federal	Pernambuco	Bahia	Alagoas
	Rio Grande do Norte	Ceará	Maranhão
	Minas Gerais	Sergipe	Paraíba
	Espírito Santo		Piauí
	Paraná		
	Santa Catarina		
	Rio Grande do Sul		
	Mato Grosso do Sul		
	Mato Grosso		
	Goiás		

5. CONSIDERAÇÕES FINAIS

A combinação do método multicritério CPP-TRI e da técnica de mineração de dados - clusterização possibilitou uma nova classificação das unidades da federação observando as similaridades entre essas unidades rompendo as fronteiras geográficas.

Dessa forma, todos os clusters puderam agrupar unidades geograficamente distantes, mas que apresentaram similaridade destacadas, tanto pela aplicação do *k-means*, quanto pela classificação do CPP-TRI, sendo assim alocadas no mesmo agrupamento.

O algoritmo *k-means* demanda como input o parâmetro k – número de clusters, sendo atribuído k=5 para que os resultados pudessem ser comparados com as macrorregiões originais. Entretanto, é possível afirmar que o valor não é adequado, uma vez que cluster C_0 ficou com apenas uma unidade.

Para a aplicação do CPP-TRI, algumas atividades foram excluídas para que a condição de classes ordenadas fosse atendida. Entretanto, os resultados apresentados foram satisfatórios e não indicam que tenham alterado significativamente os resultados. Cabe ressaltar que na presença de muitos atributos, como faz uso de probabilidades conjuntas, há uma tendência do método em convergir pra zero, sendo assim um incremento no processo de seleção de atributos pode ser necessária.

Para trabalhos futuros, sugere-se a atribuição de outros valores para k, que também impactaram nos resultados do CPP-TRI.

6. REFERÊNCIAS

[1]. Witten, I.H.; Frank, E.;Hall, M.A., *Data Mining: Practical Machine Learning Tools and Techniques: Practical Machine Learning Tools and Techniques*. 2011: Elsevier Science.

[2]. Sant'Anna, A.P.; Costa, H.G.;Pereira, V., *CPP-TRI: Um Método de Classificação Ordenada baseado em Composição Probabilística.* Relatórios de Pesquisa em Engenharia de Produção 2012. **12**(8): p. 14.

[3]. Kubrusly, L.S., *A população ocupada e a renda no Brasil: encontros e desencontros.* Economia e Sociedade, 2011. **20**(3): p. 567-600.

[4]. Sant'Anna, A.P.; Faria, F.;Costa, H.G., *Aplicação da Composição Probabilística e do Método Das K-Médias À Classificação de Municípios Quanto À Oferta de Creches.* Cadernos do IME-Série Estatística, 2013. **34**(1): p. 17.

[5]. Steiner, M.T.A.; Soma, N.Y.; Shimizu, T.; Nievola, J.C.;Neto, P.J.S., *Abordagem de um problema médico por meio do processo de KDD com ênfase à análise exploratória dos dados.* Gestão & Produção, 2006. **13**: p. 325-337.

[6]. Fayyad, U.; Piatetsky-Shapiro, G.;Smyth, P. *From Data Mining to Knowledge Discovery in Databases.* in *AAAI/MIT Press*. 1996, American Association for Artificial Intelligence.

[7]. Han, J.; Kamber, M.;Pei, J., *Data Mining: Concepts and Techniques*. 2011: Morgan Kaufmann Publishers Inc. 696.

[8]. Sant'Anna, A.P., *Aleatorização e composição de medidas de preferência.* Pesquisa Operacional, 2002. **22**: p. 87-103.

[9]. IBGE. *Censo Demografico.* 2010, Instituto Brasileiro de Geografia e Estatística.