

PROCEDIMENTOS DE DATA MINING NA DEFINIÇÃO DE VALORES PARA AS ANÁLISES DE MULTICRITÉRIOS COMO APOIO À TOMADA DE DECISÕES E ANÁLISE ESPACIAIS URBANAS

Dayan Magalhães Castro
Ana Clara Mourão Moura

Universidade Federal de Minas Gerais - UFMG
Instituto de Geociências - IGC
dayanpuc@yahoo.com.br
anaclara@yahoo.com

RESUMO

A localização geográfica é um dos fatores que influenciam a tomada de decisão para a instalação de instituições públicas, entre as quais citamos postos de saúde, assim como serviços de uso coletivo, entre os quais citamos as instituições financeiras, para as quais o local de implantação deve ser avaliado antecipadamente visando o melhor atendimento da população. A forma de avaliação proposta é o estudo da distribuição espacial de variáveis como renda média, IDH, taxa de crescimento, índice de pobreza, entre outros. Este artigo tem como objetivo mapear postos de saúde e bancos em Belo Horizonte, com vistas a caracterizar seus padrões de localização, o que dará apoio à decisão por novas instalações segundo técnicas de mineração de dados com o uso do software gratuito WEKA. Esta técnica foi usada a fim de detectar os padrões implícitos nos dados através das camadas de variáveis, utilizando-se como classificador o método de Naive Bayes. O resultado da análise da técnica de mineração foi comparada com uma análise de multicritérios utilizando o software ArcGis. Uma vez identificados os padrões espaciais que favorecem a implantação de postos de saúde e de bancos na cidade, observou-se que alguns equipamentos apresentaram comportamento espacial diferente do padrão, aleatório em relação à amostra, o que é motivo de investigação dos fatores relacionados às suas escolhas de localização. Procedimentos metodológicos como os apresentados são de grande utilidade em estudos que lidam com escolha e caracterização de principais variáveis componentes relacionadas às condições de distribuição espacial de serviços, fenômenos e potenciais urbanos e ambientais.

Palavras chaves: Mineração de dados, Análise de multicritérios, análise urbana.

ABSTRACT

The geographical location is one of the factors that influence decision making for the installation of public institutions, among which we cite health posts, as well as services for collective use, including financial institutions, for which the deployment location should be evaluated in advance in order to increase the public service quality. The evaluation of the proposal is the study the spatial distribution of variables such as average income, human development index, growth rate, poverty rate, among others. This paper explore the mapping of health centers and banks in Belo Horizonte, in order to characterize their patterns of localization, which will support the decision by new ones using techniques of data mining with the use of free software WEKA. This technique was used to detect implicit patterns in data through the layers of variables, using as classifier the method of Naive Bayes. The result of the data mining technique was compared with a multi-criteria analysis using ArcGIS software. Once identified the spatial patterns that favor the establishment of health centers and banks in the city, it was observed that some equipment had different spatial behavior than the standard, random for the sample, which is cause for investigation of factors related to their choice of location. Methodological procedures like those presented are very useful in studies dealing with choice and characterization of key variables related components to the conditions of spatial distribution of services, and phenomena of urban and environmental potential.

Keywords: Data mine, Multicriteria Analysis, urban analysis.

1 INTRODUÇÃO

Em todos os campos de pesquisa é dramático o grande acúmulo de dados provenientes de diversas fontes. De acordo com Fayyad (1996) é urgente uma nova geração de teorias computacionais e ferramentas para ajudar os homens na extração de conhecimento. No campo geográfico isso também é observado.

Uma das etapas do processo de extração de conhecimento é a chamada Mineração de Dados que é a extração automática de padrões implícitos em grandes bancos de dados, que são muito difíceis de discernir devido ao tamanho das bases e o grande número de variáveis envolvidas. Técnicas estatísticas eficientes, combinadas com teoria da informação, têm sido usadas e desenvolvidas para esse fim (White et al., 2005).

O resultado dessa seleção feita através da mineração de dados são objetos potencialmente interessantes, os quais podem ser utilizados em outras ferramentas de análise. Uma dessas ferramentas é a Análise de Multicritérios, que segundo Moura (2009) é um procedimento metodológico de cruzamento de variáveis amplamente aceito nas análises espaciais. Também conhecida como Árvores de Decisões ou como Análise Hierárquica de Pesos. O procedimento baseia-se no mapeamento de variáveis por plano de informação e na definição do grau de pertinência de cada plano de informação e de cada um de seus componentes de legenda para a construção do resultado final. A matemática empregada é a simples Média Ponderada e há pesquisadores que utilizam a lógica Fuzzy para atribuir os pesos e notas.

1.1 MINERAÇÃO DE DADOS

A utilização da mineração de dados neste estudo serviu para definir o padrão de localização das variáveis estudadas, postos de saúde e de instituições financeiras. Para a classificação foi usado o classificador probabilístico simples, naive Bayes que segundo Sala (2009), é baseado na aplicação do teorema de Bayes (inferência Bayesiana), com fortes (naive) suposições independentes. Também é conhecido como “modelo de características independentes”. Em termos simples, o classificador naive Bayes sugere que a presença (ou ausência) de uma característica particular de uma classe não tem relação com a presença (ou ausência) de qualquer outra característica. Por exemplo, uma fruta pode ser considerada uma maçã se for vermelha, redonda e com cerca de quatro polegadas de diâmetro. Embora estas características dependam da existência de outras características, o classificador naive Bayes considera que todas estas propriedades contribuem independentemente para a probabilidade desta fruta ser uma maçã (Zhang, 2004).

De acordo com Sala (2009) a matriz de confusão gerada pelo weka, em termos gerais, é uma representação em linhas e colunas, correspondendo às áreas de teste e treinamento. A matriz de confusão de uma hipótese h oferece uma medida efetiva do modelo de classificação, ao mostrar o número de classificações corretas versus as classificações preditas para cada classe, sobre um conjunto de exemplos T . Outra ferramenta valiosa para se mensurar a qualidade da classificação é o índice κ que segundo Jensen (2005), é muito utilizado para dar idéia de quanto as observações se afastam daquelas esperadas, frutos do acaso, indicando assim quão legítimas as interpretações são.

1.2 ANÁLISE DE MULTICRITÉRIOS

Segundo Moura (2007) o procedimento de análise de multicritérios é muito utilizado em geoprocessamento, pois se baseia justamente na lógica básica da construção de um SIG: seleção das principais variáveis que caracterizam um fenômeno, já realizando um recorte metodológico de simplificação da complexidade espacial; representação da realidade segundo diferentes variáveis, organizadas em camadas de informação; discretização dos planos de análise em resoluções espaciais adequadas tanto para as fontes dos dados como para os objetivos a serem alcançados; promoção da combinação das camadas de variáveis, integradas na forma de um sistema, que traduza a complexidade da realidade; finalmente, possibilidade de validação e calibração do sistema, mediante identificação e correção das relações construídas entre as variáveis mapeadas.

2 METODOLOGIA

O primeiro passo foi criar uma base de dados composta pelos dados de Belo Horizonte através da Geominas e dados do último censo levantados pelo IBGE. Foram escolhidas as seguintes informações sócio econômicas: Índice de Desenvolvimento Humano médio (IDHm), taxa de pobreza, total da renda mensal, taxa de crescimento, taxa de indigência, frequência escolar, taxa de mortalidade, número de habitantes por quilômetro quadrado e o índice Gini médio. Para os dados de infraestrutura foram utilizadas as camadas de quantidade de indústria, de comércio e de serviço. Foram usados também os limites de bairro e dos pontos das instituições financeiras e de postos de saúde da capital. Todas as camadas com exceção dos postos de saúde e dos bancos estavam agrupadas por setor censitário. Após a devida conversão de cada camada no mesmo Datum, utilizamos o SAD 69 zona 23 Sul, realizou-se então a etapa de criação das matrizes. Esse procedimento é necessário, pois segundo Moura (2009), há fortes tendências para o predomínio das operações dos modelos em formatos matriciais (*raster*). A questão se justifica pela relação de topologia implícita ao processo matricial, o que não

só otimiza o cruzamento de dados, como também é condição *sine qua non* em alguns modelos. Para a criação das matrizes a partir de cada camada sócio econômica utilizamos como padrão um pixel de 20 x 20 metros. Assim todas as matrizes continham o mesmo número de pixels (1.600, 1.100) gerados a partir de um limite geográfico único (598.000, 7.813.000 e 620.000, 7.781.000). O software ArcGis foi utilizado nesta etapa do trabalho assim como na análise de multicritérios.

Após a preparação de todas as matrizes realizou-se a etapa da extração dos valores dos pixels de cada camada, nas localizações de cada ponto. Primeiramente os postos de saúde e em seguida a extração dos valores da camada para os bancos. Conforme é apresentado na figura 1 as matrizes foram sobrepostas e foi executada a ferramenta *Sample* de extração dos valores no ArcGis, acessada em *Spatial Analyst tools* → *Extraction* → *Sample*.

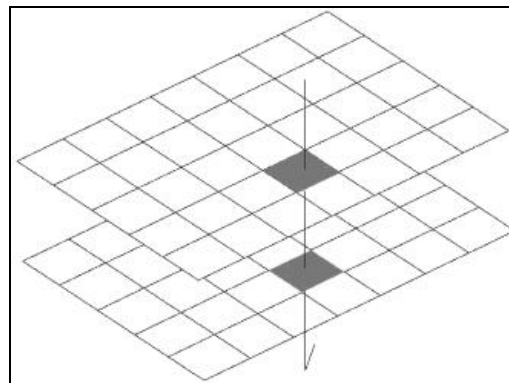


Figura 1 - Extração dos valores dos pixels em cada matriz. (Moura, 2009)

A figura 2 mostra o resultado da extração dos valores do pixel para as amostras de postos de saúde. Essa extração foi executada novamente para os bancos.

Attributes of sample_weka2												
Rowid	MASK	X	Y	Z Z Z2C1	Z Z Z2C2	Z Z Z2C3	Z Z Z2C4	Z Z Z2C5	Z Z Z2C6	Z Z Z2C7	Z Z Z2C8	Z Z Z2C9
1	0	612810,859185	7798829,272221	115	1598	1072	150	2	87	63	45	1
2	1	609065,397875	7808392,599737	198	804	721	85	2	79	48	39	0
3	2	605046,436667	7794538,570279	222	86	89	17	3	83	62	42	1
4	3	606761,933112	7794032,421499	223	347	347	39	2	82	59	44	0
5	4	612859,138779	7800959,806154	152	395	294	56	2	85	48	32	1
6	5	611112,959436	7801166,804759	150	248	214	40	1	82	64	40	0
7	6	606622,759244	7798864,285219	261	396	327	60	2	83	59	45	2
8	7	605543,316566	7793493,149344	34	183	183	42	7	78	51	38	0
9	8	604904,44271	7793244,868995	34	183	183	42	1	83	62	42	1
10	9	609623,061868	7810851,328038	201	112	71	7	7	79	53	43	2
11	10	616247,228384	7806477,70268	180	108	91	12	9	77	46	39	4
12	11	604143,727815	7808089,180415	0	273	346	36	8	80	52	41	5
13	12	605675,52267	7800136,940712	55	157	113	20	1	83	59	45	2
14	13	609277,089717	7809558,106658	197	224	208	19	2	79	48	39	0
15	14	613910,108453	7802693,999349	172	256	194	26	2	80	45	31	2
16	15	612049,278472	7796688,680298	118	5267	2099	241	1	94	100	80	0
17	16	603641,718982	7801033,271463	53	75	42	3	2	88	59	40	0
18	17	603668,542812	7796783,361364	220	49	30	1	2	85	60	39	0
19	18	606492,202315	7796992,809478	91	2369	1738	273	1	71	50	43	-1
20	19	602211,422595	7787913,065254	253	12	12	5	1	75	58	43	3
21	20	611163,127336	7805002,377717	187	213	170	21	8	82	56	37	3
22	21	611379,020819	7797069,384165	110	18913	9390	531	3	93	100	79	-1
23	22	606961,186421	7795317,614485	71	690	423	66	3	92	78	60	0
24	23	612710,036878	7794567,744445	216	1723	726	105	7	96	20	60	1
25	24	609326,523734	7797024,023236	89	3797	2763	371	2	93	100	79	-1
26	25	613451,670821	7795873,473324	120	49	9	3	3	88	67	53	0
27	26	603282,962217	7790580,381842	228	1233	1237	107	1	84	61	37	0
28	27	608630,505969	7794779,102177	74	194	129	14	1	96	58	41	0

Figura 2 - Tabela com os valores dos pixels extraídos das matrizes a partir dos pontos dos postos de saúde.

Após a extração as tabelas foram editadas e formatadas para se adequarem ao formato de entrada padrão do Weka. Inserido os dados no software de mineração de dados foi executada uma classificação utilizando o classificador *naive bayes* para essa tarefa. Não foi utilizado nenhum filtro ou préprocessamento para as amostras, apenas a classificação. Para demonstrar a qualidade da classificação apresentamos no próximo tópico os resultados da matriz de confusão e do índice kappa.

3 RESULTADOS OBTIDOS

Com o resultado inicial da mineração de dados, através do software Weka, foi possível perceber que em algumas das camadas a correlação não foi tão forte quanto em outras. A partir daí executou-se uma nova rodada no software, com as camadas que tiveram uma correlação maior. Chegando ao resultado mostrado na figura 3.



Figura 3 - Gráfico apresentando a correlação entre as camadas estudadas.

Essa segunda rodada norteou assim a análise de multicritérios, onde a partir da seleção das camadas através do estudo da mineração de dados, foi feita a análise posterior. As camadas escolhidas foram apenas as camadas de IDH, taxa de pobreza, Renda, frequência escolar e taxa de mortalidade. Essas por sua vez tinham uma correlação mais alta.

A matriz de confusão gerada a partir das variáveis estudadas na mineração de dados é mostrada na figura 4.

```

=== Confusion Matrix ===
      a  b  <-- classified as
366   0 |  a = BANCO
 16 116 |  b = Posto_de_saude
  
```

Figura 4 - Matriz de confusão das amostras

O índice kappa apresentado para estas amostras foi de 0,91, que é considerado excelente de acordo com Jensen (2005).

As camadas que se mostraram representativas nos estudos da mineração de dados foram classificadas em cinco classes, utilizando-se o seguinte padrão: nota 1 onde os valores das variáveis eram melhores, nota 10 onde os valores eram piores e notas 3, 5 e 7 para valores intermediários. Para a realização da etapa da álgebra de mapas foi utilizado o padrão de 20% no peso de cada camada, pois assim, todas as camadas teriam a mesma influência sobre o mapa final. A partir execução da álgebra percebemos uma maior distribuição de instituições financeiras ao longo das áreas com um melhor índice (figura 5). O que condiz com a realidade, onde o padrão de localização dos bancos é determinado baseando-se na distribuição dos recursos financeiros. Para os postos de saúde a localização visa alcançar uma parcela maior da população. Sendo assim, pela natureza das variáveis utilizadas, a dispersão dos postos de saúde é maior em relação aos bancos que estão mais agrupados na região central da capital.

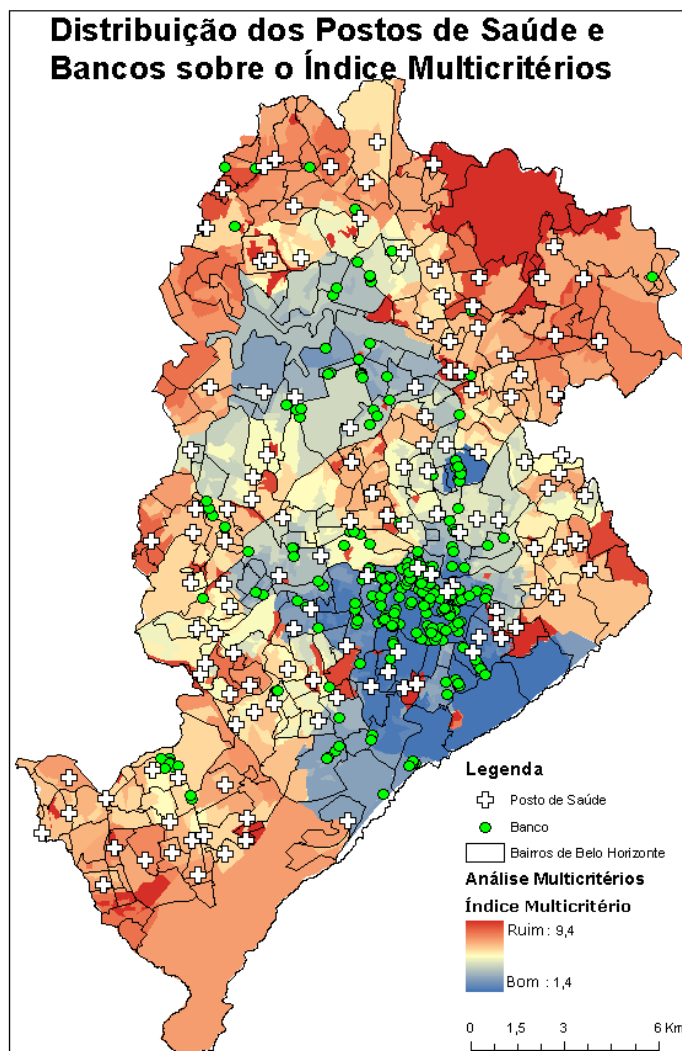


Figura 5 - Distribuição espacial dos postos de saúde e dos bancos na capital.

Como resultado da análise dos dados no software Weka apresenta-se a figura 6. Onde são mostrados os valores dos pontos do tipo “posto de saúde” que foram classificados de forma incorreta. A única amostra que geograficamente não condiz com o padrão de localização dos bancos é a amostra selecionada de azul na figura 6. Na figura 7 podemos encontrá-la ao sul do mapa. Analisando graficamente a

amostra percebe-se que a mesma está em uma região limítrofe entre valores bons e ruins. O que nos leva a duvidar da avaliação do software de mineração de dados. Uma das razões deste falso negativo pode estar relacionada com o baixo valor encontrado para a taxa de pobreza.

Attributes of Erro Classificação											
OID	INSTANCE	X	Y	PREDICTED	TIPO	TOTRENSAL	IDHMEDIO	POBREZA	FREQESC	MORTAL	Shape *
0	193	612049,28	7796688,68	BANCO	Posto	124721	94,9	80,2	121	12	Point
1	387	611859,47	7796506,57	BANCO	Posto	124721	94,9	80,2	121	12	Point
2	238	610943,03	7797232,79	BANCO	Posto	187205	93,7	79,4	100	12	Point
3	293	609326,52	7797024,02	BANCO	Posto	297029	93,7	79,4	100	12	Point
4	488	611379,02	7797069,38	BANCO	Posto	340308	93,7	79,4	100	12	Point
5	91	606961,19	7795317,61	BANCO	Posto	398565	92,1	61	104	11	Point
6	296	607520,4	7795937,71	BANCO	Posto	466078	92,1	61	104	11	Point
7	292	612884,65	7795034,21	BANCO	Posto	813602	96,5	45,4	106	7	Point
8	39	608704,7	7789177,99	BANCO	Posto	830348	94,1	39,9	105	11	Point
9	138	613611,42	7795018,81	BANCO	Posto	856052	96,5	45,4	106	7	Point
10	497	610298,85	7794576,26	BANCO	Posto	990300	96,4	68,1	110	7	Point
11	443	608630,51	7794779,1	BANCO	Posto	1083991	96,5	41,8	110	7	Point
12	394	609445,99	7793480,36	BANCO	Posto	1090922	95,2	43,1	110	11	Point
13	298	609973,59	7793927,86	BANCO	Posto	1321271	96,4	54,6	112	7	Point
14	249	612710,04	7794567,74	BANCO	Posto	1688369	96,5	60	113	6	Point

Figura 6 - Tabela com os valores das amostras classificadas de forma incorreta.

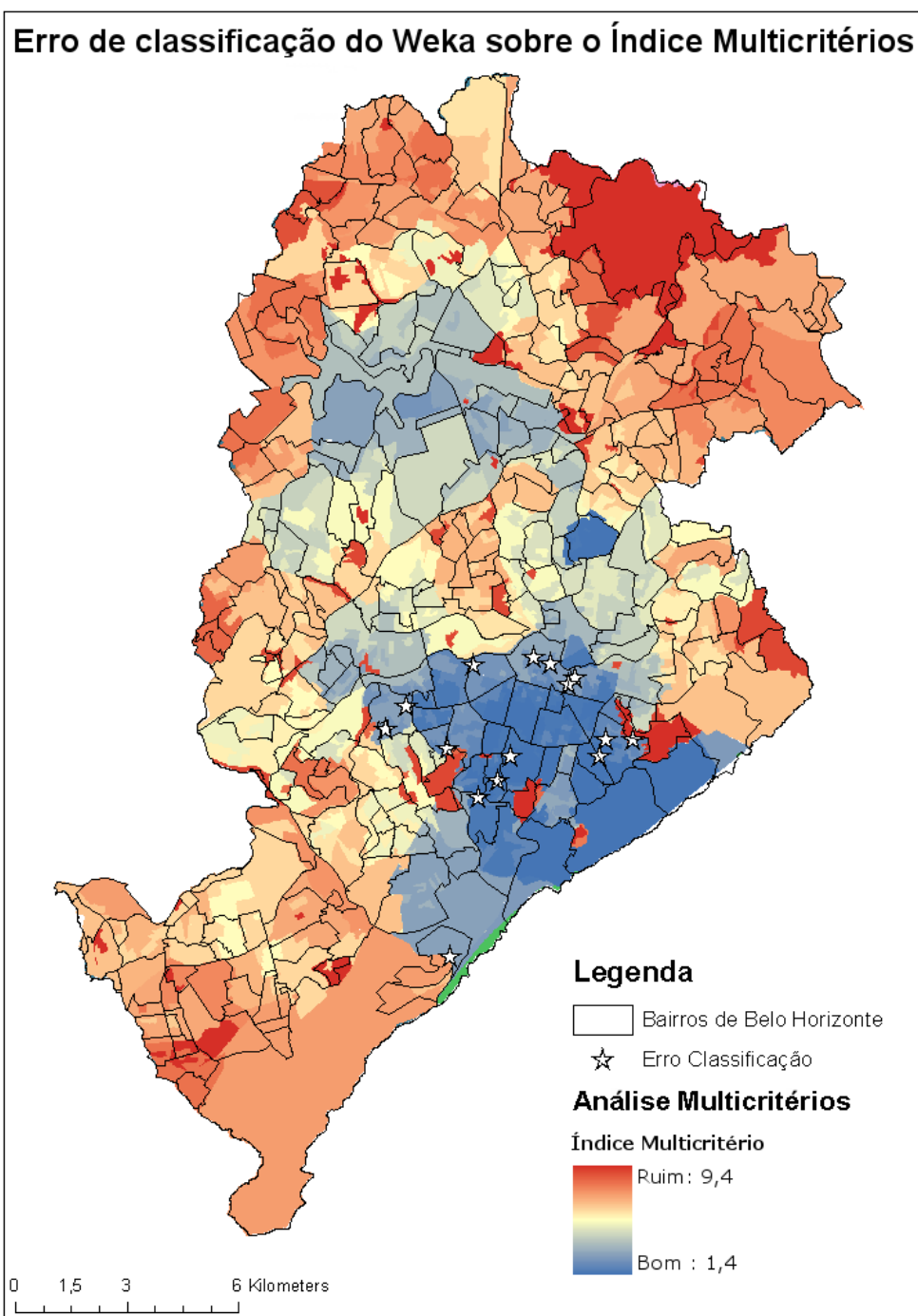


Figura 7 - Mapa com os erros de classificação sobre o Índice Multicritérios.

4 CONCLUSÃO

Através do estudo apresentado nesse artigo torna-se clara que a análise de variáveis sócio econômicas, especializadas geograficamente, podem ser estudadas através de softwares de mineração de dados. Com o objetivo de promover uma melhor avaliação em estudos posteriores, como a Análise de Multicritérios. Mesmo não avaliando a componente espacial nas variáveis trabalhadas no Weka, foram

obtidos resultados satisfatórios quanto a indicação das variáveis a serem utilizadas na álgebra de mapas, como também na definição do erro da classificação pelo método naive bayes.

A classificação através do Weka também apresentou um padrão de localização interessante nas amostras classificadas de forma incorreta. Tais amostras demonstraram que o padrão de localização dos bancos, em áreas economicamente mais ativas, é

de certa forma mais utilizado. Tendo em vista que o Weka errou na classificação de alguns dos postos de saúde da área onde o índice, apresentado na análise de multicritérios, era melhor na capital. Portanto para o padrão de localização espacial onde estavam inseridos os pontos eram apresentados como instituições financeiras.

É necessário ainda estendermos os estudos realizados aqui com o foco na amostra cuja classificação falhou, compreendendo as causas do falso negativo. Outro passo importante seria ampliar os estudos para a região metropolitana como um todo realizando assim uma análise mais macro.

5 REFERÊNCIAS BIBLIOGRÁFICAS

Fayyad, U., G. P. Shapiro, P. Smyth, 1996. From data mining to knowledge discovery in databases. AI Magazine. Volume 17(3), pp. 37-54.

Jensen, John R, 2005. introductory digital image processing: a remote sensing perspective. 3rd. ed. Pearson Prentice Hall, New Jersey, 526 páginas.

Moura, A. C. M., 2009. Discussões metodológicas para aplicação do modelo de Polígonos de Voronoi em estudos de áreas de influência fenômenos em ocupações urbanas – estudo de caso em Ouro Preto – MG. In Anais VII Encontro Nacional da Associação

Brasileira de Estudos Regionais e Urbanos - ENABER, São Paulo, Brasil, 9-11 setembro 2009, FEA/USP.

MOURA, A. C. M., 2007. Geoprocessamento no apoio a políticas do programa Vila Viva em Belo Horizonte-MG: intervenções em assentamentos urbanos precários. In Anais do XXIII Congresso Brasileiro de Cartografia, Rio de Janeiro, Brasil, 21-24 out 2007, SBC, pp. 1544-1553.

Sala, M. E., 2009. Caracterização ambiental das veredas a partir do uso da mineração de dados. In Seminário apresentado ao Programa de Pós-Graduação do Departamento de Geografia da Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brasil.

White, A. B.; P. Kumar; D. Tcheng, 2005. A data mining approach for understanding topographic control on climate-induced inter-annual vegetation variability over the United States. Remote Sensing of Environment. Volume 98, pp. 1-20.

ZHANG, H., 2004. The optimality of naïve bayes. University of Brunswick, American Association for Artificial Intelligence. Fredericton, CA, 2004.