PRIORIZAÇÃO DE AÇÕES DE PESQUISA AGROPECUÁRIA BASEADA EM MINERAÇÃO DE DADOS

Maria do Carmo Ramos Fasiaben[1]
Hércules Antonio do Prado[2]
Marcelo Fragomeni Simon[3]
Jaime Hidehiko Tsuruta[4]
José Reynaldo Ramos Machado Jr.[5]

RESUMO

Neste artigo é apresentado um método para a definição de alvos prioritários de pesquisa baseado na produção e rendimento municipais dos produtos arroz, feijão, milho, soja e trigo. Os estudos foram realizados com a utilização de ferramentas de mineração de dados, especificamente, os modelos k-médias e árvores de decisão, implementados na ferramenta *open source* Weka. Os resultados foram alocados no mapa do Brasil por meio de um sistema de informação geográfica, interpretados e discutidos. A partir do ensaio realizado verificou-se a validade do método na caracterização de grupos homogêneos de municípios. Com base na análise das características dos grupos e nos objetivos de projetos de pesquisa específicos, podem-se definir locais prioritários para pesquisa.

PALAVRAS-CHAVE: Mineração de dados, análise de agrupamentos, priorização de pesquisa.

STATING PRIORITIES FOR AGRICULTURAL RESEARCH BASED ON DATA MINING

ABSTRACT

In this paper it is presented a preliminary study that aims to formulate a method for defining targets for research. The study is based in historical data of production and productivity, at municipal level, of paddy, bean, maize, soybean, and wheat. The study was carried out with support of Weka, an open source tool. It were applied, specifically, the k-means method and decision trees. The results were plotted in the Brazil's map, by means of a geographical information system, interpreted, and discussed. On the basis of this study, it was verified how suitable is the method to characterize homogeneous groups of municipal districts. From the results of this analysis and the objectives of specific research projects, it is possible to define more adequate research efforts.

KEYWORDS: Data mining, clustering analysis, research priority.

1. INTRODUÇÃO

O contraste entre a necessidade de maiores investimentos na pesquisa agropecuária e a histórica escassez de recursos governamentais reclama o desenvolvimento de métodos para uma melhor focalização dos esforços de pesquisa. Esta focalização pode ser realizada a partir de diferentes referenciais: importância regional na produção de diferentes produtos, características agroecológicas locais que facilitem a extrapolação de resultados, logística que viabilize a instalação de ensaios e reduza as despesas com viagem, entre outros fatores. Este trabalho tem como objetivo identificar regiões prioritárias para o desenvolvimento

[1] Engenheira Agrônoma, Mestre em Economia Aplicada, Embrapa SGE, maria.ramos@embrapa.br
[2] Doutor em Ciências da Computação, Embrapa Cerrados, hercules@cpac.embrapa.br
[3] Biólogo, Mestre em Ecologia, Embrapa SGE, marcelo.simon@embrapa.br
[4] Doutor em Engenharia da Computação e Informação, Embrapa SGE, jaime.tsuruta@embrapa.br
[5] Analista de Sistemas, Embrapa Cerrados, reynaldo@cpac.embrapa.br

de pesquisas com base na distribuição geográfica da produção e rendimento e na sobreposição de diversos produtos. Desta forma, pesquisas relacionadas a diversos produtos poderiam ser realizadas em uma mesma região, estado ou município permitindo uma redução de gastos, principalmente com deslocamento de pesquisadores.

2. MATERIAL E MÉTODOS

Foram utilizados dados de produção e rendimento de 5.543 municípios do Brasil, (média dos anos 1999, 2000 e 2001), obtidos na série Produção Anual Municipal (PAM) do IBGE. Visando auxiliar o processo de interpretação dos resultados obtidos durante a análise dos dados, foi agregada ao arquivo de treinamento a zona macroagroecológica, que define, em grande parte, a aptidão agrícola de uma determinada região. Tais zonas foram definidas a partir do cruzamento de informações sobre o tipo de vegetação, relevo e características do solo (textura, drenagem e fertilidade) (EMBRAPA, 1992). As 92 zonas estabelecidas formam um gradiente de aptidão, que vai desde Preservação (zonas 1 a 26), Extrativismo (27 a 41), Pecuária (42 a 54) até Lavoura (55 a 92). A vinculação de um município a uma zona macroagroecológica foi feita a partir da comparação visual de mapas da malha municipal e das zonas macroagroecológicas (BRASIL, 1992). Como um município pode conter várias zonas, ele foi associado àquela que predomine em extensão. Especificamente, o arquivo de treinamento contém as seguintes variáveis: código do município, nome do município, unidade da federação, zona macroagroecológica, produção e rendimento de arroz, feijão, milho, soja e trigo.

Estes dados foram processados com o uso de algoritmos do sistema Weka[1], ferramenta descrita por Witten e Frank (2000), que oferece diversas opções de tratamento de dados, incluindo famílias de técnicas de análise de agrupamentos e classificação. Neste trabalho foram construídos modelos de agrupamentos com base no algoritmo k-médias e de classificação com base em árvores de decisão. Descrevemos, sucintamente, estas duas técnicas a seguir. O método k-médias deriva o seu nome do fato de iniciar com um conjunto de k "sementes" escolhidas como pontos de partida para os centróides. A partir da escolha dos valores iniciais dos centróides, cada registro de entrada é associado ao centróide mais próximo. Após a associação de todos os registros, cada centróide é recalculado como a média de todos os registros a ele associados. A atualização dos centróides é repetida até que eles não se modifiquem mais. Uma árvore de decisão é um grafo que relaciona variáveis (representadas por nós) por meio de seus valores (representados por arcos no grafo). A partir de um nó (*pai*), é feito um teste para decidir qual nó *filho* deve ser pesquisado a seguir. Inicialmente, foram construídas quatro configurações de agrupamentos, utilizando-se o algoritmo k-médias, ora com os dados de rendimento, ora com os dados de produção, variando em cada caso o parâmetro S igual 10 e igual a 30. O parâmetro S é utilizado para definir um conjunto aleatório de pontos. O número de grupos (parâmetro n do Weka) foi fixado, arbitrariamente, em quatro. Uma análise preliminar dos resultados apontou para um conjunto mais interessante de grupos com S=30. Em seguida, foi gerado um conjunto de grupos a partir dos dados de rendimento e produção, com S=30, obtendo-se uma separação melhor do que aqueles gerados com a utilização dos rendimentos ou da produção isoladamente. Para auxiliar a interpretação dos agrupamentos, foi gerada uma árvore de decisão utilizando-se unicamente a zona macroagroecológica. A partir desta abordagem pode-se identificar as características de aptidão agrícola de cada agrupamento. Para facilitar a visualização, os municípios foram mapeados de acordo com o resultado da análise de agrupamento.

[1] Weka é um software de código aberto distribuído conforme a *GNU General Public License.* http://www.cs.waikato.ac.nz/ml/weka/index.html

3. RESULTADOS

A análise de rendimentos e produções permitiu a caracterização dos grupos mostrada na Tabela 1 e mapeada conforme a Figura 1:

Grupo A: Dos municípios do país, 13% foram alocados no grupo A. Neste grupo estão contidos 43% dos municípios do Centro-Oeste e 17% do Sudeste, que representam 90% dos membros deste grupo. O grupo A se caracteriza por apresentar médias de rendimento superiores às nacionais para todas as culturas analisadas, com exceção do trigo. Os municípios deste grupo são responsáveis por 44% da produção nacional de soja, 32% de arroz, 26% de milho e 14% de feijão. Os municípios que compõem o grupo A estão presentes com maior expressão nas zonas macroagroecológicas de número 60, 61, 62 e 76, sendo as duas primeiras importantes produtoras de milho e soja. Todas estas zonas se caracterizam por boa aptidão para lavoura.

Grupo B: O grupo B abarca 44% dos municípios do país, chamando a atenção o fato de que 94% dos municípios da Região Nordeste e 71% dos da Região Norte estão nele contidos. Os municípios destas duas grandes regiões representam 82% do agrupamento. Também o integram municípios do Sudeste e Pantanal. Este grupo tem como característica os baixos níveis de produção e produtividade em todas as culturas. Entretanto, neste grupo se produziu 31% do feijão nacional no período analisado, 11% do arroz e 6% do milho. A produção de soja e trigo é inexpressiva neste grupo. Municípios sem vocação agrícola também podem ser aqui alocados. Elementos do grupo B estão distribuídos por diferentes zonas macroagroecológicas, predominantemente naquelas com aptidão para preservação, extrativismo e pecuária.

TABELA 1: Distribuição dos municípios por grupo e região.

Região	Grupo A			Grupo B			Grupo C			Grupo D		
	Municípios	% Região	% do Grupo	Municípios	% Região	% do Grupo	Municípios	% Região	% do Grupo	Municípios	% Região	% do Grupo
N	34	8%	5%	316	71%	13%	0	0%	0%	97	22%	6%
NE	37	2%	5%	1689	94%	69%	0	0%	0%	64	4%	4%
CO	200	43%	28%	48	10%	2%	36	8%	5%	178	39%	11%
SE	284	17%	39%	346	21%	14%	38	2%	6%	987	60%	58%
S	169	14%	23%	53	4%	2%	599	50%	89%	368	31%	22%
Total	724			2452			673			1694		

Grupo C: Cerca de 12% dos municípios do país foram incluídos no grupo C. Este é o grupo que tem a maior concentração de municípios da Região Sul (50%), que por sua vez representam 89% do total do grupo, sendo os restantes distribuídos entre as regiões Sudeste e Centro-Oeste. Aqui não aparecem municípios das Regiões Norte e Nordeste. Este grupo tem como principal característica os altos rendimentos obtidos com todas as culturas, inclusive trigo. Observam-se aqui, também, os maiores rendimentos médios de arroz, o que também indicaria bom nível de utilização de tecnologia moderna. No grupo C concentra-se a maior parte da produção nacional de trigo (97%), soja (54%), milho (52%) e feijão (36%), além de 22% da produção de arroz. Os municípios do grupo C estão distribuídos de forma preponderante nas zonas 67, 61, 87, 92, 70, 75, 77 e 54, que apresentam aptidão para lavoura.

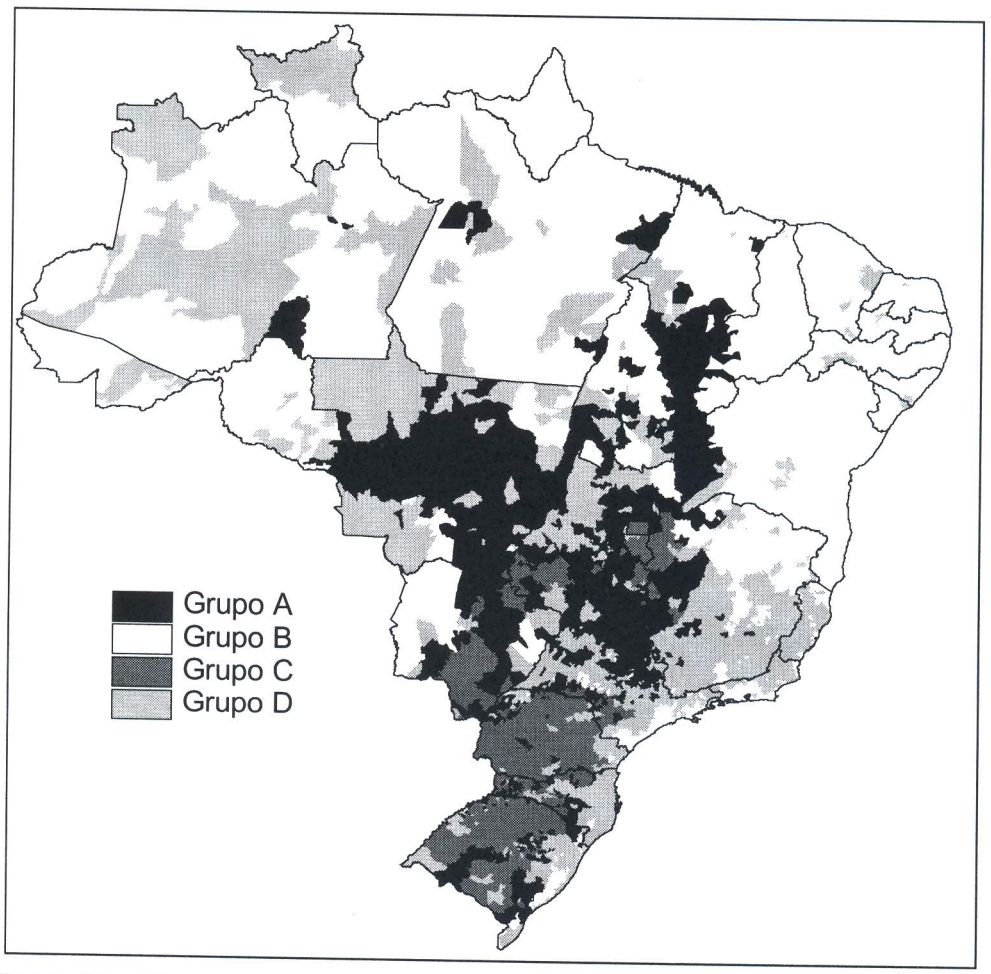

FIGURA 1: Classificação dos municípios do Brasil de acordo com análise de agrupamento (k-médias), baseada na produção e rendimento municipais de arroz, feijão, milho, soja e trigo (média do triênio 1999 a 2001).

Grupo D: Foram alocados 31% do total de municípios do país no grupo D. Aqui estão 60% dos municípios da Região Sudeste (que representam 58% do grupo), além de municípios das Regiões Sul, Centro-Oeste, Norte e Nordeste. Trata-se de um grupo que reúne municípios com rendimentos acima da média nacional para as culturas do arroz, feijão e milho, não apresentando, entretanto, relevância nas culturas de soja e trigo. Cerca de 35% do arroz nacional é produzido por municípios deste grupo, assim como 19% do feijão e 16% do milho. Os municípios deste grupo estão localizados de forma preponderante nas seguintes zonas macroagroecológicas: 68, 72, 73, 74, 90, 50, 51, 53, 59, 23, 13, com diferentes classes de aptidão.

4. DISCUSSÃO E CONCLUSÃO

Este trabalho constitui-se num ensaio preliminar de estabelecimento de agrupamentos para orientar ações de P&D. Outros métodos podem ser testados, assim como outras variáveis que orientem a formação de grupos, para seleção dos que apresentem resultados mais satisfatórios aos objetivos propostos, além de maior simplicidade. Neste sentido, a ferramenta Weka mostrou-se muito útil, uma vez que traz diversas opções de

tratamentos de dados com interfaces de fácil utilização. No caso específico deste trabalho, a utilização das variáveis produção e rendimento permitiu que fosse diminuído o efeito do tamanho dos municípios na análise, o que não aconteceu quando se utilizaram somente dados referentes à produção. A listagem dos grupos por zona macroagroecológica permite visualizar a importância relativa desses grupamentos nas zonas. Assim, se o objetivo do trabalho de P&D for estender tecnologias já consagradas a um maior número de produtores representativos de diferentes grupos, pode-se optar pela escolha de zonas que concentrem o maior número de grupos em seu interior, como por exemplo a zona de número 92 ou a de número 61.

Se, por outro lado, o objetivo for desenvolver pesquisas em zonas de agricultura menos desenvolvida, como por exemplo para atingir os objetivos do Projeto Fome Zero, pode-se optar pela zona 43, onde predominam municípios do grupo B, com produções significativas de milho e feijão, mas com baixos rendimentos. Dentro de cada zona, então, pode-se escolher o município que atenda às condições logísticas mais satisfatórias.

Como se trabalhou em escala nacional, os resultados são bastante genéricos. A aplicação de estudos deste tipo com caráter regionalizado é interessante, assim como a inclusão de outros produtos. Estes resultados preliminares mostram que estudos deste gênero são interessantes na orientação das ações institucionais.

5. REFERÊNCIAS BIBLIOGRAFIA

BRASIL. Ministério da Agricultura e Reforma Agrária. Delineamento macroagroecológico do Brasil-1992/93. Osasco: MARA: EMBRAPA-SNLCS: Geograft Didática, 1992.

EMBRAPA. Secretaria de Administração Estratégica. Regionalização (delineamentos) macroagroecológico do Brasil. Brasília, DF, 1992. 122p.

WITTEN, I. H.; FRANK, E. Data mining: practical machine learning tools and techniques with Java implementations. San Francisco: Morgan Kaufmann Pub., 2000. 371p.