UTILIZAÇÃO DE REGRAS DE ASSOCIAÇÃO PARA RELACIONAR DADOS METEOROLÓGICOS E DADOS DE PRODUTIVIDADE DO CAFÉ NO ESTADO DE SÃO PAULO

Celso Macedo Junior[1]; Priscila Pereira Coltri [2]; Hilton Silveira Pinto[3]; Jurandir Zullo Junior[3]

[1]Mestrando, Faculdade de Engenharia Agrícola (FEAGRI) e CEPAGRI/UNICAMP, Campinas-SP, celso@cpa.unicamp.br
[2]Doutoranda, Faculdade de Engenharia Agrícola (FEAGRI). e CEPAGRI/UNICAMP, Campinas-SP, pcoltri@cpa.unicamp.br
[3]Pesquisador, Dr., CEPAGRI/UNICAMP, Campinas-SP, {jurandir, hilton} @cpa.unicamp.br,

RESUMO: O estado de São Paulo é o terceiro maior produtor de café do Brasil, sendo que este produto agrícola representa o terceiro item da economia de exportação do país. Apesar de tamanha importância comercial, o setor cafeeiro ainda sofre com perdas principalmente devido a elementos meteorológicos. Excesso de temperatura na fase do florescimento pode gerar o abortamento das flores. Em contrapartida, bom aporte hídrico nesta fase pode auxiliar no sucesso da colheita da cultura. Sendo assim, o presente estudo teve como objetivo produzir regras de associação entre os dados meteorológicos – precipitação e temperatura máxima – médios dos meses iniciais da primavera – setembro e outubro - utilizando os algoritmos *Apriori*, *Predictive Apriori* e *Tertius* do *software* de mineração de dados WEKA. O algoritmo *Apriori* não obteve resultados satisfatórios, pois não gerou regras com suporte mínimo acima de 0.01. Já os algoritmos Predictive Apriori e Tertius, produziram regras que demonstram a importância do incremento de precipitação na fase de florescimento para o aumento de produtividade anual. Além disso, também mostraram que os melhores valores de produtividade anual encontravam-se no intervalo de temperatura máxima que estava compreendido entre [23ºC – 27ºC].

Palavras-chave*:* regras de associação, dados meteorológicos, produtividade do café, São Paulo.

USE OF THE ASSOCIATION RULES TO RELATE METEOROLOGICAL DATA AND COFFEE PRODUCTIVITY IN THE STATE OF SAO PAULO

ABSTRACT: The state of São Paulo is the third largest producer of coffee in Brazil, and this product represents the third item in the export economy of the country. Despite such commercial importance, the coffee sector still suffers a loss mainly due to meteorological factors. Excess temperature in the flowering stage can generate the abortion of the flowers. However, good water supply at this stage may help in the success of the harvest of the crop. Thus, this study aimed to produce rules of the association between average meteorological data - precipitation and maximum temperature – of the early spring months - September and October - using the algorithms *Apriori*, *Predictive Apriori* and *Tertius* of data mining software WEKA . The algorithm *Apriori* not obtained satisfactory results, because no rules created with support less than 0.01. But the algorithms *Predictive Apriori* and *Tertius* produced rules that demonstrate the importance of the increase of precipitation during the flowering to the annual increase in productivity. Further, it showed that the highest values of annual productivity were at the maximum temperature range that was between [23ºC – 27ºC].

Key words: association rules, meteorological data, coffee productivity, Sao Paulo

INTRODUÇÃO

A importância do café para o Brasil é indubitável, uma vez que se trata do principal produto agrícola brasileiro de exportação. Desta forma, esse produto agrega um considerável volume de recursos à balança comercial. Como exemplo disso, em 2000, as exportações de café alcançaram cerca de 1,6 bilhões de dólares, e em termos de produção, a Região Sudeste do Brasil concentra a maior produção nacional de café (IBGE, 2000). O estado de São Paulo é o terceiro maior produtor de café atual, só ficando atrás de Minas Gerais e Espírito Santo (CONAB, 2008).
Existem duas espécies de café cultivadas no mundo: *Coffea Arábica* ou, simplesmente, café Arábica, e a *Coffea Canephora*, o café Robusta ou *Conillon*. Porém, no estado de São Paulo especificamente é produzido predominantemente a primeira espécie. O café arábica é originário de áreas tropicais da Etiópia localizadas entre 6° e 9°Norte de latitude, em altitudes que variam entre1.600 e 2.000 m. A temperatura média anual nesta região é de 18ºC a 20°C (mínima de 4ºC a 5°C e máxima de 30 a 31°C) e a precipitação anual é de 1.500 a 1.800 mm (CAMARGO & PEREIRA, 1994).

Eventos climáticos desfavoráveis à cafeicultura podem induzir perdas significativas o ano todo, já que esta é uma cultura perene. Problemas como geadas, deficiências hídricas, ventos frios constantes, alta variabilidade espacial de chuvas são algumas das adversidades climáticas que o café sofre. No entanto, problemas com altas temperaturas

são também objeto de preocupação para os cafeicultores. Entre 18ºC e 22ºC, estão as temperaturas médias anuais do ar mais favoráveis ao cultivo do café arábica, estando ideal entre 19ºC e 21ºC, desde que sejam regiões livres ou pouco sujeita a geadas. As regiões que possuem temperatura média anual inferior a 18ºC e superior a 23ºC são consideradas inaptas para o café arábica (CAMARGO, 1985; ASSAD et al., 2001).
Segundo CAMARGO (1985) e THOMAZIELLO *et al.* (2000), as temperaturas do ar elevadas na fase de florescimento poderão dificultar o êxito das floradas e provocar a formação de "estrelinhas", ou seja, o abortamento de flores. Estas flores abortadas ocorrem principalmente devido à freqüência de temperaturas máximas superiores a 34ºC, e desta forma, causando perda de produtividade (PINTO *et al*., 2001). Entretanto, quando se observa a precipitação neste período, os especialistas atestam que o café produzido no cerrado produz grãos de alta qualidade porque durante a época da florada dos cafezais, as chuvas são abundantes, permitindo a brotação dos frutos (SIMÃO, 1999)

O ciclo fenológico da cafeicultura tropical brasileira é bem definido, sendo o florescimento na primavera, frutificação no verão, maturação no outono e colheita no inverno. Determinar qual são as temperaturas médias máximas no início da fase chuvosa (setembro e outubro), quando o florescimento das flores do café está tendo início, e relacioná-las com dados de suas respectivas produtividades pode acarretar um ganho de conhecimento bastante significativo. No entanto, dados meteorológicos de chuva e temperatura possuem alta variabilidade espacial e temporal, dados não-balanceados, presença de dados faltantes e ruídos, além do grande número de instâncias.

A mineração de dados tem mostrado uma excelente ferramenta para predições agroclimáticas na agricultura. Trabalhos como o de BUCENE *et al*. (2002) demonstram como essa técnica de descoberta de conhecimento pode auxiliar em alertas de geadas e deficiência hídrica, especificamente no caso do café. Entretanto, há poucos trabalhos no domínio da agricultura utilizando técnicas de "clusterização" ou associação. Algoritmos de associação vêm sendo desenvolvidos para tentar identificar padrões em dados históricos. AGRAWAL (1994) propôs o modelo *Apriori* que gera regras de associação em cima do suporte e confiança dos dados. Anos depois, SCHEFFER *et al.* (2001) contribui com o modelo *Predictive Apriori*, que busca uma relação entre suporte e confiança que possa maximizar a chance de uma correta predição de dados não analisados. GILLMEISTER & CAZELLA (2001) compararam os dois algoritmos de associação acima citados, além do algoritmo *Tertius* proposto por FLACH (2001), utilizando dados de indústria automotiva, demonstrando a necessidade de mais estudos com diferentes conjuntos de dados.

Sendo o café uma cultura de grande importância para o agronegócio brasileiro e sabendo que sua produtividade pode ser comprometida por eventos climáticos, foi desenvolvido o presente trabalho. Este tem como objetivo criar regras de associação entre as médias de setembro e outubro das variáveis meteorológicas estudadas – precipitação e temperatura máxima – e os valores de produtividade anual para cada município produtor de café do estado de São Paulo.

MATERIAL E MÉTODOS

Dados

Foram utilizados dados médios de precipitação e temperatura máxima dos meses de setembro e outubro fornecidos pelo IAC (Instituto Agronômico de Campinas), CPTEC (Centro de Previsão de Tempo e Estudos Climáticos), INMET (Instituto Nacional de Meteorologia), UNICAMP e ESALQ/USP. Os anos disponíveis dos dados meteorológicos variaram de acordo com a estação de cada localidade. Os dados de produtividade foram extraídos da série de 1990 a 2006 do SIDRA/IBGE.

Preparação Preliminar dos Dados

Primeiramente, foram calculadas as médias de temperatura máxima e precipitação dos dados diários dos meses de setembro e outubro para cada ano do respectivo município. Depois, foi montado um arquivo no Excel™, agregando os dados de produtividade anual municipal. Os dados meteorológicos e de produtividade anuais faltantes foram ignorados para não interferir nas regras de associação que seriam geradas posteriormente.
Além disso, houve vários municípios os quais os dados de produtividade de anos consecutivos estavam replicados. A fim de eliminar qualquer erro por preenchimento ou inserção de ruídos, foi mantida apenas uma instância.

Portanto, depois deste tratamento inicial do conjunto de dados e a análise de *boxplot* que será comentada a seguir, o número de cidades distintas do conjunto de dados foi de 93 e o de instâncias, 523, ilustradas na figura 1, que demonstra uma boa espacialização destas.

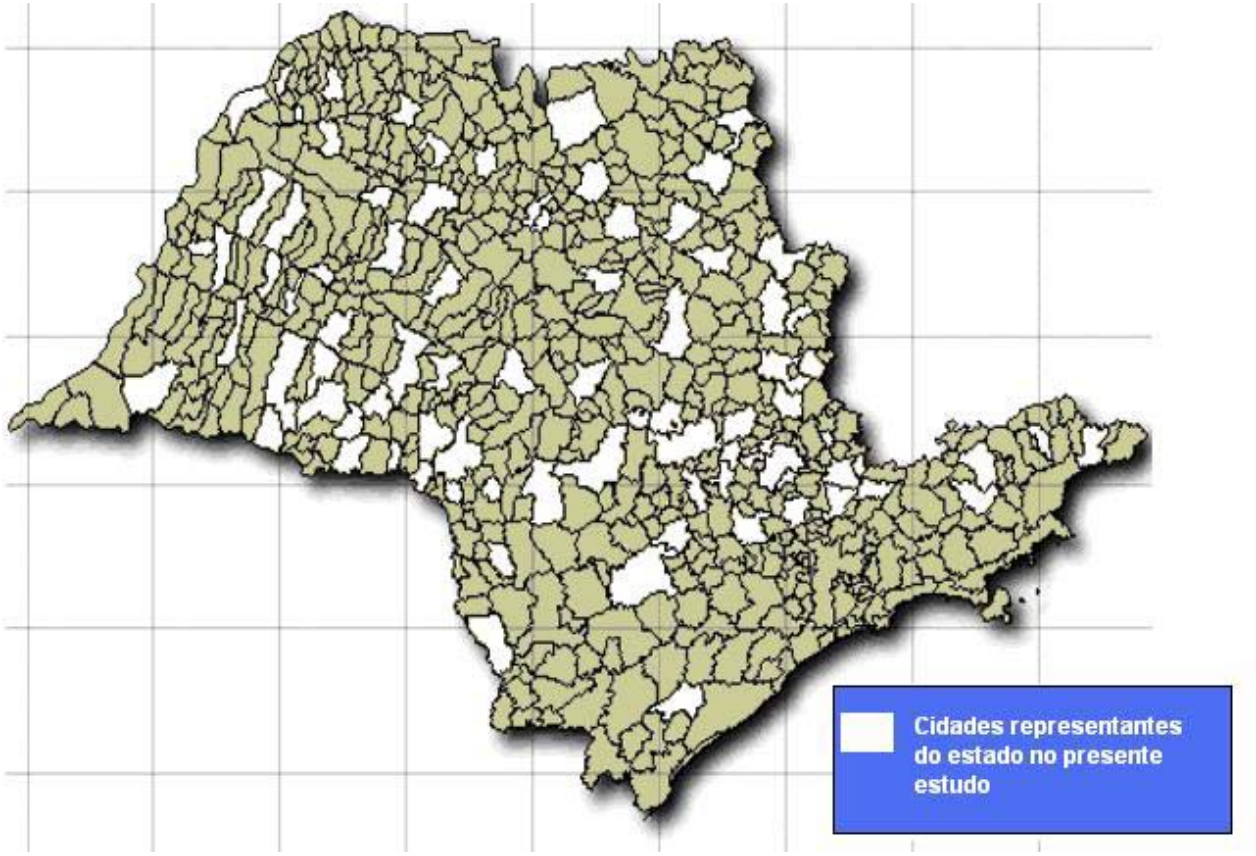

Figura 1 – Cidades representantes da produtividade do café no estado de São Paulo para o presente estudo.

Preparação dos Dados

Como já foi antecipado na sessão anterior, a fim de eliminar os possíveis *outliers* do conjunto de dados foi feito um *boxplot* com os dados ordenados e calculados os respectivos limite superior e inferior para os três atributos: temperatura máxima, precipitação e produtividade, dados pelas equações abaixo:

Limite Inferior = $Q_1 – 1,5*(Q_3 – Q_1)$ Eq. (1)
Limite Superior = $Q_1 + 1,5*(Q_3 – Q_1)$ Eq. (2)

Onde:
Q_1 = Primeiro quartil;
Q_3 = Terceiro quartil.

Para eliminar que ruídos dominassem sobre cada classe, e também por ser uma necessidade dos algoritmos de associação, foi feita a discretização pelo particionamento baseado na freqüência (*Equi-Depth*) dos atributos. Esse método apresenta resultados bastante satisfatórios para atributos numéricos, pois evita pesos diferentes entre as distintas faixas discretizadas.
O conjunto de dados possui desbalanceamento em seu número para cada cidade. Porém, como o objetivo do trabalho foi buscar regras de associação gerais para o estado, cada instância foi mantida para garantir uma maior distribuição espacial e temporal dos dados.

Algoritmos de Mineração dos Dados

Os algoritmos utilizados para gerar as regras de associação foram o *Apriori*, *Predictive Apriori* e o *Tertius* do *software* WEKA™.

RESULTADOS E DISCUSSÃO

Mineração de Dados

Apresentado a metodologia utilizada no presente trabalho, faremos uma avaliação dos resultados obtidos e quais inferências podem se inserir sobre estes.

Utilizando o algoritmo *Apriori*, não foram encontradas regras com o suporte mínimo de 0.01. Mesmo utilizando os dados sem a exclusão dos *outliers*, as regras encontradas foram redundantes e bastante gerais com suporte mínimo de 0.01. Portanto, para este conjunto de dados o uso deste algoritmo não foi satisfatório.
Para os algoritmos *Predictive Apriori* e *Tertius* foram criadas inúmeras regras de associação, porém as 5 primeiras regras foram bastante razoáveis como podem ser observadas nas tabelas 1 e 2 que estão alocadas abaixo.

Tabela 1 – Melhores regras de Associação encontradas pelo algoritmo *Predictive Apriori*

```
1.  tmax ='(27.25-27.95]'  precipitacao ='(2.25-2.65]' 5 ==> Produtividade='(1090.5-1256]' 3    acc:(0.38745)
2.  tmax ='(-inf-27.25]'  precipitacao ='(2.95-3.25]' 4 ==> Produtividade='(808-953]' 2    acc:(0.31)
3.  tmax ='(-inf-27.25]'  precipitacao ='(3.25-3.65]' 4 ==> Produtividade='(1090.5-1256]' 2    acc:(0.31)
4.  tmax ='(27.25-27.95]'  precipitacao ='(1.95-2.25]' 4 ==> Produtividade='(953-1090.5]' 2    acc:(0.31)
5.  tmax ='(-inf-27.25]'  precipitacao ='(4.75-inf)' 19 ==> Produtividade='(1637.5-2112.5]' 6    acc:(0.2763)
```

Tabela 2 – Regras de Associação encontradas pelo algoritmo *Tertius*

```
1. /* 0,143395 0,051625 */  precipitacao  = '(4.75-inf)' ==> Produtividade = '(1408.5-1637.5]' or  tmax  = '(-inf-27.25]'
2. /* 0,138321 0,059273 */  tmax  = '(-inf-27.25]' ==>  precipitacao  = '(4.75-inf)' or Produtividade = '(953-1090.5]'
3. /* 0,129162 0,061185 */  tmax  = '(-inf-27.25]' ==>  precipitacao  = '(4.75-inf)' or Produtividade = '(2112.5-inf)'
4. /* 0,128251 0,061185 */  tmax  = '(-inf-27.25]' ==>  precipitacao  = '(4.75-inf)' or Produtividade = '(808-953]'
5. /* 0,120290 0,065010 */  precipitacao  = '(4.75-inf)' ==>  tmax  = '(-inf-27.25]'
```

A tabela 1, que mostra as 5 melhores regras de associação encontradas pelo algoritmo *Predictive Apriori*, explicita uma interessante relação entre as variáveis meteorológicas – temperatura máxima e precipitação médias dos meses de setembro e outubro – com as respectivas produtividades daquele determinado ano. Interessante foi observar, independente de quais faixas as regras de associação estavam compreendidas, o quão a precipitação média do início da primavera estava relacionada com as variações de produtividade. Para as regras 2, 3 e 5, ambas as temperaturas máximas estavam compreendidas no intervalo [11.2ºC – 27.3ºC], entretanto para essa faixa de temperatura máxima média dos meses de setembro e outubro, quando há incremento na quantidade de precipitação há aumento de produtividade. Além disso, a regra de maior acurácia pode ser descrita como:

R1 – Tmax = [27 – 28] Precipitação = [2 - 3] => Produtividade [1090 – 1256] Eq. (3)

Os dados de produtividade são anuais e os dados meteorológicos são as médias dos dois meses iniciais da primavera, logo, outros fatores podem afetar a produtividade durante o ano. Mas, estas regras fornecem uma heurística para tentar compreender como o início da estação chuvosa, atrelado ao aumento de temperatura que ocorre na primavera, podem deixar indícios na produtividade anual – já que a época de florescimento é bastante problemática para a cultura do cafeeiro.

No caso da tabela 2, que mostra as 5 melhores regras de associação encontradas pelo algoritmo *Tertius*, reitera a importância da precipitação na época da florada para um bom desempenho do café, pois a regra melhor elencada foi:

R2 – Precipitação [4.8 – 6.6] => Produtividade [1408 – 1638] Eq. (4)

Além disso, as regras 2, 3 e 4 demonstram, sem analisar as precipitações, as diversas faixas de produtividade atreladas à temperatura máxima dos dois meses no intervalo entre [23ºC – 27ºC], o que mostra a importância desta faixa para ter-se uma boa produtividade da cultura estudada.
Analisando os excertos acima, podemos inferir que o algoritmo *Apriori* não foi satisfatório para este banco de dados, já que seu suporte mínimo para a criação de regras de associação ficou abaixo de 0.01. Entretanto, os algoritmos *Predictive Apriori* e *Tertius* produziram resultados razoáveis de acordo com o problema e domínio propostos. Ambos os algoritmos demonstraram a importância da precipitação no período do florescimento do café como já era esperado. Todavia, o presente trabalho quantificou para o estado de São Paulo essa quantidade média de precipitação do início da primavera, concomitantemente com os valores de temperatura máxima média deste mesmo período.

CONCLUSÕES

Neste trabalho foram desenvolvidas regras de associação entre dados meteorológicos – precipitação e temperatura média – médios dos meses de setembro e outubro, além de dados de produtividades anuais para os municípios produtores de café presentes na figura 1. Para gerar essas regras foram utilizados os algoritmos *Apriori*, *Predictive Apriori* e *Tertius*.

O algoritmo *Apriori* não gerou nenhuma regra de associação com suporte mínimo acima de 0.01, o que o tornou insatisfatório para esse conjunto de dados. Porém, os algoritmos *Predictive Apriori* e *Tertius* produziram resultados razoáveis para o complexo domínio do estudo abordado, sendo suas respectivas regras, as equações 3 e 4. Em ambas as regras podemos observar a importância do incremento da quantidade de precipitação para o melhor desempenho da cultura do café de acordo com SIMÃO (1999). Além disso, as 5 melhores regras de cada algoritmo delimitam uma faixa de temperatura máxima para o período de florescimento que contém as produtividades mais altas dos municípios do estado de São Paulo. Esse intervalo encontrado ficou compreendido entre [23.2ºC – 27.9ºC].

Outrossim, foi que este estudo pôde utilizar a técnica de regras de associação, tão pouco utilizada na agricultura, e também avaliar quais critérios de preparação dos dados podem produzir melhores resultados em dados de alta variabilidade temporal e espacial, como é o caso dos dados agroclimáticos. Para proposta futuras, talvez devamos incorporar a bianulidade produtividade do café e o maior número de dados meteorológicos.

REFERÊNCIAS BIBLIOGRÁFICAS

AGRAWAL, R. et al. Fast algorithms for mining association rules in large databases. In: **International Conference on Very Large Data Bases**, VLDB, 20., 1994, Santiago. p. 478-499.

ASSAD, E. D. et al. Zoneamento agroclimático para a cultura de café (Coffe arabica L.) no estado de Goiás e sudoeste do estado da Bahia. **Revista Brasileira de Agrometeorologia**, Passo Fundo, v.9, n.3, p.510-518, 2001.

BUCENE, L. C.; RODRIGUES, Luiz H. A.; MEIRA, Carlos Alberto Alves; Mineração de Dados Climáticos para Previsão de Geada e Deficiência Hídrica para as Culturas do Café e Cana-de-Açúcar para o Estado de São Paulo. **SBI-Agro**, Vol. 1, pp.1-5, Porto Seguro, BA, Brasil, 2003

CAMARGO, A. P. O clima e a cafeicultura no Brasil. **Informe Agropecuário**, Belo Horizonte, n126, p. 13-26, 1985.

CAMARGO, A.P.; PEREIRA, A.R. Agrometeorology of the coffee crop. **World Meteorological Organization. Geneva: WMO/TD**, 1994. n. 615, 43 p.

CONAB. Avaliação da Safra Agrícola Cafeeira de 2008. **Boletim técnico do mês de agosto**. 17p., 2008.

FLACH, P.; LACHICHE, N. Confirmation-Guided Discovery of First-Order Rules with Tertius. **In: Machine Learning**, v. 42, Issue 1/2, Jan. 2001, Kluwer Academic, USA, pages. 61-95.

GILLMEISTER, P. R. G.; CAZELLA, S. C. Uma análise comparativa de algoritmos de regras de associação: minerando dados da indústria automotiva. **In: Escola Reginonal de Banco de Dados**, 2007, Caxias do Sul. Escola Reginonal de Banco de Dados, 2007.

IBGE. Aspectos das atividades agropecuárias e extração vegetal, seção 3, p.23-61. In IBGE (ed.), **Anuário estatístico do Brasil,** 2000.

PINTO, H.S.; ZULLO JUNIOR, J.; ASSAD, E.D.; BRUNINI, O.; ALFONSI, R.R.; CORAL, G. Zoneamento de riscos climáticos para cafeicultura do Estado de São Paulo. **Revista Brasileira de Agrometeorologia**, v.9, p.495-500, 2001.

SCHEFFER, T. et al. (2001) Finding association rules that trade support optimally against confidence. **In: PKDD 2001: principles of data mining and knowledge discovery, European conference on principles of data mining and knowledge discovery,** N. 5, 20011973, v. 2168, pages. 424-435

SIMÃO, M.L.R. **Caracterização espacial da produção cafeeira de Minas Gerais: um estudo exploratório utilizando técnicas de análise espacial e de estatística multivariada.** Dissertação Mestrado. Belo Horizonte: PUC-MG, 1999. 246p.

THOMAZIELO, R. A.; FAZUOLI, L. C.; PEZZOPANE, J. R. M.; FAHL, J. I.; CARELLI, M. L. C. Café arábica: Cultura e técnicas de produção. Campinas: **Instituto Agronômico**, 2000. 82 p.