Agrupamento das regiões homogêneas de precipitação no Rio Grande do Sul pela análise de cluster e detecção de anos relativamente secos e chuvosos nas sub-regiões desse estado

David Ferreira Junior[1]
Ana Maria Heuminski De Ávila[2]

A precipitação pluvial é um elemento meteorológico altamente variável no tempo e no espaço e é a principal variável meteorológica responsável pelo sucesso ou fracasso da produção agrícola nas regiões Tropicais. O Estado do Rio Grande do Sul apresenta alta variabilidade das chuvas, sendo que aprender novas informações acerca do comportamento espaço-temporal dessa variável meteorológica é fundamental para que um adequado planejamento agrícola seja feito. Os parâmetros climáticos exercem influência sobre todos os estágios da cadeia de produção agrícola, incluindo a preparação da terra, a semeadura, o crescimento dos cultivos, a colheita, a armazenagem, o transporte e a comercialização, sendo que em regiões tropicais o principal deles é a precipitação. Entretanto, é necessário que esta seja distribuída no tempo e no espaço para que haja boa produtividade. As técnicas disponíveis em mineração de dados são ferramentas que têm se mostrado adequadas para identificar séries de dados com características semelhantes de precipitação e, desta forma, identificar regiões homogêneas. Os objetivos desse trabalho consistiram em transformar as séries históricas de dados de precipitação em zonas pluviometricamente homogêneas por meio de técnicas de agrupamento (Data Mining), aplicar a Técnica dos Quantis, visando detectar a ocorrência de anos relativamente secos ou chuvosos em cada sub-região homogênea com relação às médias dos totais

[1] Universidade Estadual de Campinas (Unicamp)
[2] Centro de Pesquisas Meteorológicas e Climáticas Aplicadas à Agricultura (Cepagri)

anuais de precipitação e relacionar a produtividade de soja do Rio Grande do Sul com o volume de chuva no período do verão (maior necessidade de água) e da colheita (em que grandes volumes de chuva tendem a prejudicar a produção). O conjunto de dados consiste em séries históricas do período de 1981 a 2010 de precipitação pluvial registradas por várias estações meteorológicas espalhadas pelo território do Rio Grande do Sul. O banco de dados utilizado foi o da Agência Nacional de Águas (ANA) com dados diários de precipitação. Os dados de produtividade da soja para o Rio Grande do Sul foram obtidos da página da Empresa de Assistência Técnica e Extensão Rural (Emater) e do Instituto Brasileiro de Geografia e Estatística (IBGE). No entanto, existiam meses com nenhum dia registrado. Inicialmente foi feita uma seleção das estações que continham registros na janela temporal de interesse (1981 a 2010). O preenchimento dos dados faltantes foi feito por meio da técnica de imputar valores pela média ponderada de 10 estações meteorológicas mais próximas (vizinhos mais próximos) utilizando valores de precipitação estimados pelo satélite *Tropical Rainfall Measuring Mission* (TRMM) e dados de estações meteorológicas de superfície disponível. Do conjunto de estações virtuais e de superfície disponíveis foi calculado o peso de cada uma das estações a partir do inverso do quadrado da distância e finalmente o cálculo da média ponderada. A correlação (correlação de Pearson) entre valores mensais estimados e verdadeiros foi de 0.77. Com os dados preenchidos, foi aplicada a técnica de agrupamento (Análise de Cluster) para detectar regiões pluviometricamente homogêneas. Foi utilizado o algoritmo *k-means*, em que os dados são agrupados de acordo com a similaridade medida pela métrica de distância euclidiana. O número ótimo de clusters foi sugerido pela análise dos gráficos da coesão interna e coesão externa (que mostra o quanto a sub-região é homogênea e o quanto as sub-regiões são heterogêneas entre si), e é o ponto em que se pode notar um "cotovelo" nesses gráficos. Esse número ótimo foi validado por especialista da área e também por justificativas descobertas nos dados. Após a identificação das zonas pluviometricamente homogêneas foi aplicada a técnica dos quantis para estabelecer os limiares de quantidade de chuva, com o intuito de classificar a ocorrência de eventos extremos anuais. O número ótimo de clusters sugerido foi 5. Segundo especialistas de conhecimento da área de estudo, o que faz sentido para o Rio Grande do Sul é 4 sub-regiões, uma vez que as diferenças esperadas são com respeito as 4 regiões geográficas mais fáceis de se perceber mudanças (regiões norte, sul, leste e oeste). A divisão encontrada pelo algoritmo *k-means* usado nesse trabalho ao supor

5 sub-regiões separa a região oeste em noroeste e sudoeste. Essa separação é diferente do que se esperava, mas realmente faz sentido. A divisão encontrada pelo algoritmo ao supor 5 sub-regiões separando a região oeste em noroeste e sudoeste é perceptível nas médias das precipitações anuais, em que a região noroeste segue geralmente a mesma tendência da região norte, tendo contudo, na maioria dos anos, um volume de precipitação um pouco menor em média. A região sul é a região mais constante em termos de precipitação ao longo do ano em média. A região litorânea passa do outono ao inverno com um aumento da precipitação, enquanto que a região sudoeste passa de um outono mais chuvoso para o inverno mais seco das sub-regiões. Depois, optou-se por investigar especificamente as chuvas no norte do estado, por ser essa a região com a maior produtividade de soja do Rio Grande do Sul. O ano 1982 obteve um volume de chuvas em janeiro/fevereiro em média ligeiramente menor que em 1986, porém observa-se que houve uma grande queda na produtividade média em 1986. Um dos fatores de forte influência nesse caso de 1986 foi o grande volume de chuvas na colheita. No ano de 1991 observa-se uma grande queda no volume de chuvas no verão, que refletiu sendo a segunda média mais baixa de produtividade para o período. O ano mais chuvoso foi o de 2002 para todas as sub-regiões, precedido por 2001 com um volume normal de chuva e tendo 2003 um volume normal também. No entanto, 2001 e 2003 registraram uma produção em média muito maior que 2002, o que é coerente, uma vez que os anos de 2001 e 2003 apresentaram maiores volumes de chuva no verão e menores volume de chuva na colheita em comparação com 2003. A partir desse estudo, conclui-se que a técnica de preenchimento de dados de precipitação pluvial por "vizinhos mais próximos" é eficiente para esse conjunto de dados; que técnica de agrupamento é eficiente na detecção de diferenças de comportamento de precipitação pluvial para o Estado do Rio Grande do Sul, além de obter uma informação nova ao separar a parte oeste do Rio Grande do Sul em duas sub-regiões; que a produtividade do Rio Grande do Sul apresenta uma tendência de crescimento linear ao longo dos 30 anos de estudo (que é coerente com o desenvolvimento e aperfeiçoamento das técnicas agrícolas); que a análise exploratória (isto é, sem formalidade estatística) dos dados mostra que os anos considerados secos no período de janeiro/fevereiro (verão) e/ou chuvosos em março/abril (colheita) no norte do Rio Grande do Sul apresentaram quedas na produtividade, mas não em todos os casos. Assim, embora não seja uma constante, a análise dos dados mostra que realmente existe certo impacto na produtividade o fato de ocor-

rerem secas no verão e/ou chuvas na colheita. O projeto foi concluído, mas algumas sugestões surgem como motivação direta, como relacionar a série histórica de produção agrícola de mais produtos do Rio Grande do Sul com outras variáveis como temperatura e altitude além da precipitação pluvial e desenvolver ou aplicar alguma técnica de preenchimento de dados faltantes para precipitação levando em conta variáveis como altitude e temperatura além da distância geográfica como é o caso da técnica de “vizinhos mais próximos”.

Palavras chave: Séries históricas, precipitação, técnica de agrupamento, produtividade agrícola, dados faltantes.