ISSN: 2317-0840

Análise da expansão da soja : Uma Aplicação de Análise Multivariada e Redes Neurais Artificiais

Marina S. Paiva[1†], Marcelo S. Nagano[2], Kuang Hongyu[3]

[1] *USP.*
[2] *USP. E-mail: drnagano@usp.br.*
[3] *UFMT. E-mail: prof.kuang@gmail.com.*

Resumo: *O estado do Mato Grosso é o maior produtor e exportador de soja no Brasil, este trabalho apresenta uma caracterização do crescimento da soja no estado de Mato Grosso, entre os anos 1990–2015. O uso concomitante de análise multivariada - Análise de Componentes Principais (PCA) e análise de redes neurais de Kohonen (SOM) podem permitir a verificação e desenvolvimento de modelos e metodologias que permitam uma melhora significativa deste tipo de estudo. Será realizada uma identificação, caracterização e análise das variáveis que tiveram contribuição (ou não) na expansão da soja no estado. A utilização em conjunto da análise componentes principais (PCA) e da análise com as redes neurais de Kohonen podem possibilitar a comprovação e o desenvolvimento de modelos e metodologias que possibilitem uma melhora significativa deste tipo de estudo e conclui-se que as uniões das ferramentas obtiveram os resultados esperados e agregaram valores entre elas, diminuindo a dimensionalidade do banco possibilitando melhores interpretações e visualização dos dados em estudo.*

Palavras-chave: Soja; Componentes principais; Redes Neurais Artificiais; Multivariada ; Mapas auto organizáveis de Kohonen.

Abstract: *The state of Mato Grosso is the largest producer and exporter of soybeans in Brazil, this work presents a characterization of soybean growth in the state of Mato Grosso between 1990 and 2015. The concomitant use of multivariate analysis - Principal Component Analysis (PCA) and analysis of Kohonen neural networks (SOM) may allow the verification and development of models and methodologies that allow a significant improvement of this type of study. An identification, characterization and analysis of the variables that have contributed (or not) to soybean expansion in the state will be performed. The joint use of the main components analysis (PCA) and the analysis with the Kohonen neural networks can enable the verification and development of models and methodologies that allow a significant improvement of this type of study and it is concluded that the tool unions obtained the expected results and added values ??between them, reducing the dimensionality of the data base allowing better interpretations and visualization of the data under study.*

Keywords: Soybean; Principal Component; Artificial Neural Network; Multivariate; Kohonen self organizing map.

[†]Autora correspondente: marinaspaiva@usp.br.

Introdução

A percepção sobre a cadeia produtiva da soja se tornou realidade no Brasil, no final da década de 60, quando se começou a perceber que a soja poderia se tornar um produto de grande valor comercial, devido a intensa necessidade de produção de farelo para criação suínos e aves. Neste contexto, torna-se necessário o desenvolvimento não só de estudos de cadeia produtiva, mas também da sua prática na agroindústria. A história do estado do Mato Grosso está diretamente relacionada ao desenvolvimento do agronegócio. A expansão do preço da soja no mercado mundial em meados de 1970 despertou interesse dos produtores agrícolas. Além disso, o governo brasileiro também passou a aplicar incentivos de recursos para a cultura da soja (FEARNS, 2011).

Estrategicamente, o Brasil se beneficia de vantagens competitivas em relação aos outros países produtores de soja, uma vez que: O escoamento da safra brasileira ocorre na entressafra americana, quando os preços atingem as maiores cotações no mercado mundial. O país então, passou a investir em tecnologia para adaptação da cultura às condições brasileiras, processo liderado pela Empresa Brasileira de Pesquisa Agropecuária (EMBRAPA). Os investimentos em pesquisa genética levaram à "tropicalização" da soja, permitindo que o grão fosse plantado com sucesso em regiões de baixas latitudes, mais especificamente entre o trópico de capricórnio e a linha do equador (ZOCKUN, 1978).

Neste artigo é realizada uma análise sobre a expansão territorial e produtiva de soja no Estado de Mato Grosso usando o método de Estatística Multivariada, especificamente análise de componentes principais (PCA) e Redes Neurais Artificiais utilizando redes auto organizáveis de Kohonen (SOM). Os resultados são provenientes dos dados do Instituto Brasileiro de Geografia e Estatística (IBGE), através do Sistema SIDRA referente a PAM (Pesquisa Agrícola Municipal), do USDA (United States Department of Agriculture) e do INPE (INPE - Instituto Nacional de Pesquisas Espaciais) do período entre 1990 e 2015. Quando o interesse é verificar como as amostras se relacionam, ou seja, o quanto estas são semelhantes, segundo as variáveis utilizadas na pesquisa, destacam-se dois métodos de estatística multivariada: a análise de agrupamento e a análise fatorial com análise de componentes principais. A técnica denominada de análise de componentes principais, conhecida popularmente por PCA, foi desenvolvida por Karl Pearson em 1901 com base no artigo de Hotelling de 1933. Seu objetivo principal é o de explicar a estrutura da variância e da covariância de um vetor aleatório, composto por p-variáveis aleatórias, através da construção de combinações lineares das variáveis originais. Estas combinações lineares são chamadas de componentes principais e são não-correlacionadas entre si (HONGYU et al., 2015). Se tivermos p-variáveis originais é possível obter p componentes principais. No entanto, é desejável diminuir o número de variáveis a serem avaliadas, ou seja, a informação contida nas p-variáveis originais é substituída pela informação contida em k (k ¡ p) componentes principais não-correlacionadas. Com isso, o sistema de variabilidade do vetor aleatório composto das p-variáveis originais é aproximado pelo sistema e pode ser medida através da avaliação da proporção de variância total explicada por elas.

Redes neurais artificiais são modelos computacionais inspirados no sistema nervoso de seres vivos (SILVA et al., 2010). A a estrutura das redes neurais foi desenvolvida a partir de modelos de sistemas nervosos biológicos e do cérebro humano. As redes neurais artificiais podem ser empregadas em diversos problemas, tais como: aproximação universal de funções; controle de processos; reconhecimento ou classificação de padrões; agrupamento de dados; sistemas de previsão; otimização de sistemas; memórias associativas (SILVA et

al., 2010). Devido sua grande aplicabilidade, diversas redes foram desenvolvidas.

Neste artigo, será utilizada a rede de Kohonen. Esta rede é formada por uma rede neural de camada única. Os dados são apresentados à entrada e os neurônios de saída são organizados com uma estrutura de vizinhança simples. Cada neurônio está associado a uma referência no vetor (o vetor de peso), e cada ponto de dados é mapeado em um neurônio junto com o mais próximo (distância euclidiana)(KOHONEN, 1982; KOHONEN, 1989; KOHONEN, 1991).

No processo de execução do algoritmo, cada ponto de dados funciona como uma amostra que direciona o movimento dos vetores de referência para o valor dos dados desta amostra. Os vetores associados são alterados durante o processo de aprendizagem e tendem para os valores da distribuição dos dados de entrada. O valor característico de um cluster pode ser intuitivamente entendido como o típico valor dos dados no cluster (ASTEL et al., 2007).

No final do processo o conjunto de dados de entrada é particionado em conjuntos disjuntos (os clusters) e o peso associado a cada neurônio é o valor característico do cluster associado ao neurônio no caso unidimensional, que é o caso desse trabalho. Com isso o cluster das partições é mais fácil de visualizar e não é difícil comparar o comportamento dos dados nos aglomerados correspondentes.

O objetivo deste trabalho é avaliar e identificar o quanto a soja se expandiu no período de 1990 à 2015 no estado de Mato Grosso. Identificar, caracterizar e analisar as variáveis que tiveram contribuições (ou não) na expansão da soja no estado de Mato Grosso, sendo as variáveis de estudo: área plantada, área colhida, quantidade produzida, valor da produção, média da quantidade de chuva, taxa de desmatamento, produtividade e média do preço mundial. Também será observado se as variáveis escolhidas para estudo têm correlação entre elas, ou seja, se há ligação na expansão desta no estado de Mato Grosso e a distribuição destas variáveis por meio de clusters que expressam similaridades entre as variáveis. Após essa análise, os resultados serão utilizados para avaliar a importância da união das ferramentas aplicadas.

O restante do artigo está organizado como segue: Na Seção 2 será visto o problema proposto, na Seção 3 os resultados encontrados e as análises dos resultados e na Seção 4 são apresentados conclusões e futuros trabalhos.

Problema Proposto

O agronegócio brasileiro é uma atividade próspera, segura e rentável. Atualmente, o agronegócio é o cargo chefe na economia brasileira e é um dos setores que mais gera emprego e renda no país. No contexto mundial atual, o Brasil situa-se como o celeiro do Agronegócio, contém com a ajuda de vários planos de incentivo do governo e alto investimento em modernas tecnologias, fazendo deste um setor eficiente e competitivo no mercado mundial (HIRAKURI e LAZZAROTTO, 2011).

Em vista do objetivo geral deste artigo, será feita a identificação, caracterização e análise das variáveis que tiveram aumento (ou não) na expansão da soja. Por meio da análise multivariada, será feita uma avaliação de componentes principais para melhor visualização da correlação e explicação entre as variáveis. É proposto também uma rede neural de Kohonen, a qual sua principal característica é não ser supervisionada. Essa rede foi proposta no estudo devido à baixa dimensão da rede e sua estrutura simples. Propõem-se representar de forma simples, os aglomerados por meio de vetores associados

a cada neurônio e mapear a topologia do conjunto de dados de entrada. Em relação à seleção geográfica, para que a pesquisa fosse consistente e se pudesse realmente avaliar o crescimento da soja, foram coletados dados de todos os municípios e somados, gerando um único valor denominado Mato Grosso.

Resultados de implementação e análise

Análise de Componentes Principais

Para a obtenção dos resultados, foram utilizados os softwares R para programação das análises multivariadas e Visual Studio com programação em C++ para implementação da rede de Kohonen. Após os resultados da rede serem obtidos, foi utilizado o Toolbox do Matlab para a obtenção dos gráficos. O uso da análise de componentes principais (PCA) teve como objetivo reduzir a dimensão do conjunto de variáveis e facilitar a interpretação da independência entre estas. Para que isso ocorra, foram obtidas as combinações lineares das variáveis originais que, geometricamente representam a seleção de novos sistemas de coordenadas obtidos pela rotação do sistema original que tem p variáveis aleatórias como eixos das coordenadas. As desvantagens de se utilizar essa técnica é a redução das variáveis que acarretam na redução da variabilidade das informações. Outro caso é que ela nem sempre funciona eficientemente, mesmo reduzindo sua quantidade de variáveis,esse é o caso quando as componentes principais são as próprias variáveis originais, sendo pouco correlacionadas. O primeiro componente principal se define como a combinação linear das variáveis originais que tem variância máxima. Os Valores deste primeiro componente de n indivíduos se representam por vetor $Z1$ dado por: $Z1 = \mathbf{Xa}$.

Como as variáveis tem média zero, $Z1$ também terá média nula, então sua variância será:

$$\frac{1}{n} z1' z1 = \frac{1}{n} a1' \mathbf{X'Xa1} = \mathbf{a1'Sa_1}, \tag{1}$$

em que S é a matriz de variância e covariância das observações.Pode-se maximizar a variância aumentando o módulo do vetor $\mathbf{a1}$. Para que a maximização da equação acima tenha solução, deve-se impor uma restrição ao módulo do vetor $\mathbf{a1}$. Neste artigo, estabelece-se o valor $\mathbf{a'1a1 = 1}$. Com isso, essa restrição é introduzida mediante o multiplicador de Lagrange:

$$M = a_1' S a_1 - \lambda(a_1' a_1 - 1). \tag{2}$$

Esta expressão é maximizada da forma habitual derivando respectivamente os componentes de a_1 igualando a zero, então teremos:

$$\mathbf{Sa_1} = \lambda a_1. \tag{3}$$

Que quer dizer que a_1 é um vetor próprio da matriz S e λ, o seu valor correspondente. Para determinar o valor próprio valor de S, multiplica-se a esquerda da equação 3 por a_1:

$$a_1' S a_1 = \lambda a_1' a_1 = \lambda. \tag{4}$$

Tabela 1: Correlação componente com variáveis.

Variável	CP1	CP2
X1	0,9937	0,0035
X2	0,9985	0,0050
X3	0,9981	0,0297
X4	-0,0533	0,9986
X5	0,9972	0,0152

Tabela 2: Composição componentes principais e variâncias.

Componentes	Autovalor	Variância (%)	Variância ac. (%)
1	3,9779	79,5584	79,5584
2	1,0000	19,9661	99,5245
3	0,0168	0,3357	99,8602
4	0,0066	0,1324	99,9926
5	0,0004	0,0074	100,0000

Para a determinação do número de componentes principais (CP), verificou-se que, como os dois primeiros CP's gerados a partir desta análise tem autovalores maiores que 1 ($\lambda = 1$) (RENCHER, 2002) e foi responsável por 99,52% da variância total no conjunto de dados, os dois CP's foram retidos, com o auxílio do screeplot 1.

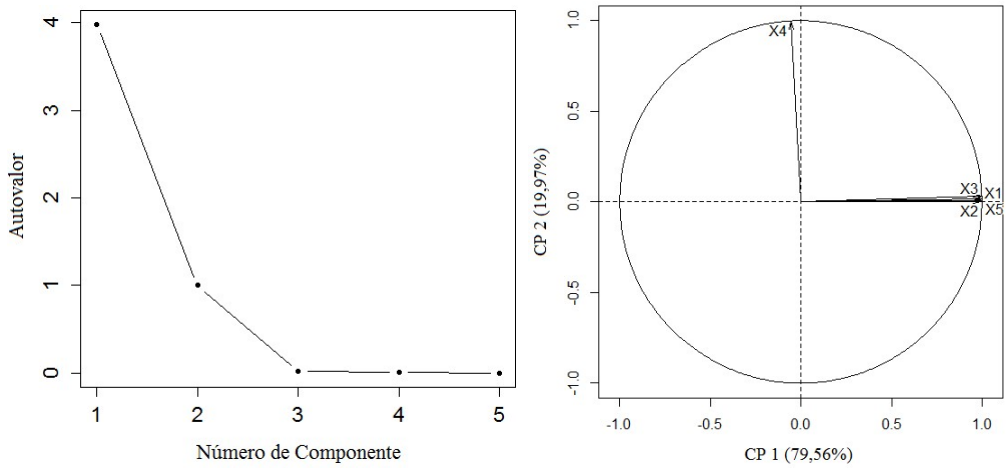

Figura 1: Scree Plot e Biplot CP1 X CP2.

Portanto, dois primeiros CP's resumem efetivamente a variância amostral total e podem ser utilizados para o estudo do conjunto de dados. Também na 1 mostra-se o Biplot CP1xCP2 com as cidades sobre a produção da soja. Pode-se concluir que, de acordo com os dados referentes à expansão da soja no estado do Mato Grosso e com a ACP, *Sorriso, Nova Mutum, Campo Novo do Parecis, Querência, Nova Ubiratã e Sapezal* possuem maiores taxas de produção de soja no estado e principalmente sobre maiores números de área colhida e valor da produção de CP1. As cidades apresentadas no cluster 1 foram às cidades que apresentaram menores valores na área colhida, valor da produção e taxa de desmatamento.

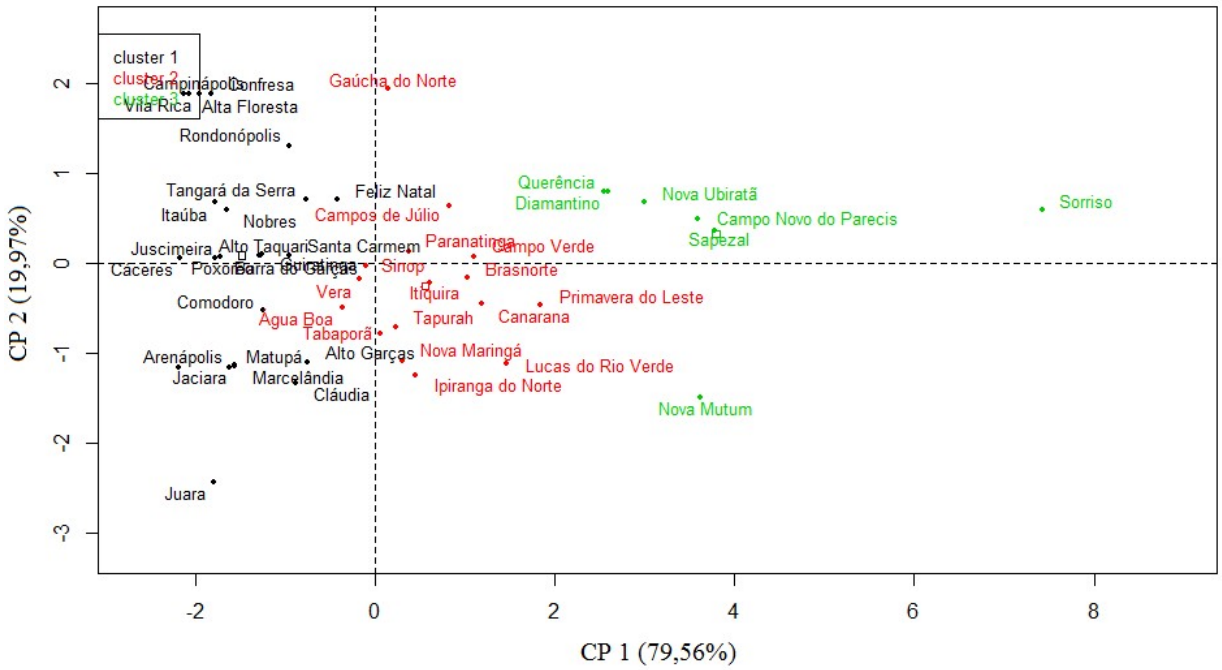

Figura 2: Biplot CP1 X CP2 com indivíduos.

Mapas Auto Organizáveis de Kohonen

A rede neural ou mapa de atributos auto organizados de Kohonen, possui um conjunto de elementos de entrada (cujo número coincide com a dimensão dos vetores que compõem o espaço de fator) e um conjunto de elementos de saída correspondentes a clusters. Os elementos de entrada são projetados para distribuir o vetor de entrada entre os elementos de saída da rede. Os valores de peso dos clusters podem ser interpretados como valores de coordenadas que descrevem a posição do cluster no espaço de dados de entrada (KHAYKIN,2008). Observa-se que é aconselhável organizar os clusters na forma de uma rede bidimensional, porque essa topologia assegura que cada neurônio tenha muitos vizinhos. Este arranjo determina quais elementos serão ajustados dentro de um raio do vencedor. O princípio básico para o processo de aprendizado competitivo é a concorrência entre os neurônios, pois se tem como objetivo sair como vencedor desta. Ressalta-se que o processo é não-supervisionado (não há saída desejada). Para o ganhador, será feito um ajuste dos pesos, proporcional aos valores apresentados como dados de entrada, no sentido de melhorar seu desempenho para a próxima competição. A métrica utilizada para a proximidade entre os dois pontos é :

$$distj^{(k)} = \sqrt{\sum_{i=1}^{n}(X_i{}^{(k)} - X_I{}^{(k)})^2} \qquad \text{onde } j = 1, ..., n_1$$

A $distj^{(k)}$ quantifica a distância (norma euclidiana) entre o vetor de entrada representando a k-ésima amostra, em relação aos vetores de pesos do j-ésimo neurônio. Logo, o neurônio j que obtiver a menor distância será declarado o vencedor da competição frente a amostra k e receberá ajustes nos vetores de pesos de maneira que fique cada vez mais

perto da amostra. Pode-se elencar então três aspectos relevantes para a configuração de um SOM são:

- Definição da organização espacial dos neurônios;
- Delimitação dos conjuntos de vizinhança de cada neurônio;
- Detalhamento do critério de ajuste do vetor de pesos do neurônio vencedor e seus respectivos vizinhos

As instruções computacionais do algoritmo que detalham a fase de treinamento de mapas auto-organizáveis de Kohonen são apresentados na sequência:

Algoritmo Kohonen - Fase de Treinamento

Início {

1 Definir o mapa topológico da rede;

2 Montar os conjuntos de vizinhança $[\Omega^k]$;

3 Iniciar o vetor de pesos de cada neurônio considerando os valores das n_1 primeiras amostras do treinamento;

4 Obter o conjunto de amostras de treinamento $[\mathbf{x}^k]$;

5 Normalizar os dados (vetores) e pesos;

6 Especificar a taxa de aprendizagem $[\eta]$;

7 Iniciar contador de épocas;

8 Repetir as instruções;

8.1 Em todas as amostras de treinamento $[\mathbf{x}^k]$ fazer :

i Calcular as distâncias euclidianas entre $[\mathbf{x}^k]$ e $[\mathbf{w}^j]$ conforme a expressão acima;

ii Declarar o neurônio vencedor, com menor distância euclidiana;

iii Ajustar vetor de pesos do vencedor;

iv Ajustar o vetor de pesos do neurônio vizinho;

v Normalizar o vetor de pesos que foi ajustado na instrução anterior;

8.2 Época $\leftarrow +1$

Enquanto não houver mudanças significativas nos vetores de pesos;

9 Analisar o mapa visando extração das características;

10 Identificar as regiões que possibilitam a definição das classes;

} Fim

Fonte: Silva et al., 2010

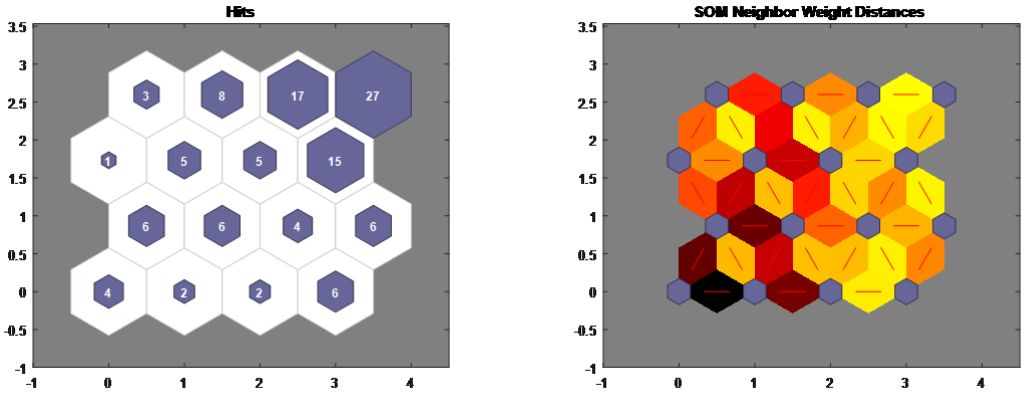

Figura 3: Gráfico de Hit e Gráfico SOM dos pesos das distâncias dos vizinhos.

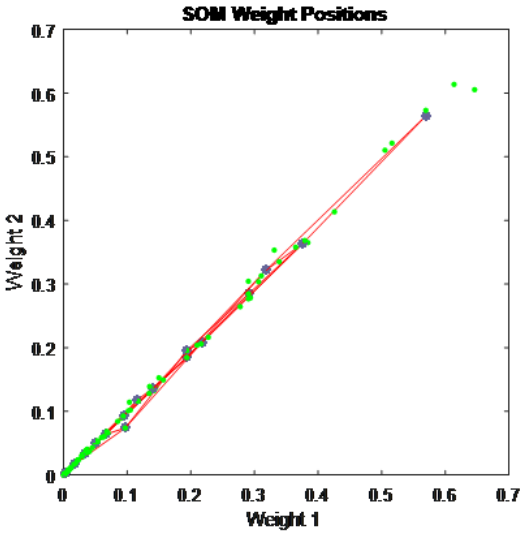

Figura 4: Gráfico SOM dos pesos das posições dos vetores. Fonte: Autor

Após gerar os resultados por meio do Visual Studio, foi utilizado o *software* MatLab para a geração dos gráficos representativos da rede de Kohonen. Foram utilizados 117 dados, sendo separados 100 para treinamento e 17 para teste.

Para a criação dos gráficos, foi selecionada a configuração (4 X 4). Os 16 vetores que formam o conjunto de testes não levam em conta as especificidades dos atributos de entrada e, portanto, essa escolha de exemplos de treinamento não deve ser considerada ideal para maximizar a entropia.

A Fig. 3 mostra a distribuição do espaço de 16 clusters, construídos pela rede de Kohonen (tabela retangular de 4 x 4). Os números nas células correspondem ao número de elementos nos clusters. A Fig. 3 (direita) também mostra os neurônios como manchas cinza-azuladas e suas relações de vizinho direto com linhas vermelhas.

As manchas vizinhas são coloridas de preto a amarelo para mostrar o quão próximo o vetor de peso de cada neurônio está para seus vizinhos. Com isso, observa-se uma maior

aproximação dos vetores em destacados de amarelo localizados à direita do gráfico, onde se encontra a maior localização dos dados no cluster.

Na Figura 4 são apresentados os vetores de entrada como pontos verdes e mostra como o SOM classifica o espaço de entrada mostrando pontos cinza-azulados para o vetor de peso de cada neurônio e conectando neurônios vizinhos com linhas vermelhas, logo podemos observar que os pontos de entradas estão mais concentrados em uma área, porém com variações nas conexões com os vizinhos.

Conclusão

Neste artigo, verificou-se a importância destas análises para as tomadas de decisões, na dinâmica do processo de avaliação do desempenho da soja no mercado do agronegócio.

Nota-se pelo presente estudo que o crescimento da produtividade se deve, principalmente pela ampliação da área plantada. Pode-se também observar que todas as variáveis escolhidas para estudo obtiveram uma alta correlação.

Em contrapartida, a taxa de desmatamento não foi explicada com o aumento da área plantada. Conclui-se que as uniões das ferramentas obtiveram os resultados esperados e agregaram valores entre elas, possibilitando melhores interpretações e visualização dos agrupamentos.

A utilização concomitante da análise multivariada com a análise de componentes principais e da análise com as redes neurais de Kohonen pode possibilitar a comprovação e o desenvolvimento de modelos e metodologias que possibilitem uma melhora significativa deste tipo de estudo. Especialistas se mostram otimistas quanto aos resultados obtidos pela utilização da associação do SOM com alguma outra técnica estatística (ASTEL et al., 2007), já havia feito uma comparação entre a aplicação de SOM para classificação de conjuntos de dados muito grandes com as análises tradicionais como Análise de Agrupamentos *cluster analysis* e PCA.

Como sugestão para pesquisas futuras, propomos um estudo mais abrangente do assunto, com mais variáveis e outras técnicas de multivariada para que outros tipos de informações que possam ser extraídos destes.

Referências bibliográficas

ASTEL, A.; TSAKOVSKI, S.; BARBIERI, P.; SIMEONOV, V. Comparison of self organizing maps classification approach with cluster and principal components abalysis for large environmental data sets. *Water Research*, v.41, n.19, p.4566-4578, 2007.

MOHAMED, E.; ABDELATIFI, E.; EL MOUTAOUAKIL, K. Learning algorithm of kohonen network with selection phase. *WSEAS Transactions on Computers*, v.11, n.11, p.387-396, 2012.

FEARNSIDE, P. M. Soybean cultivation as a threat to the environment in Brazil. *Environmental Conservation*, v.28, n.1, p.23-38, 2001.

HIRAKURI, M.H.; LAZZAROTTO, J.J. *Evolução e perspectivas de desempenho econômico associadas com a produção de soja nos contextos mundial e brasileiro.* 3 ed.,

Embrapa Soja, Londrina, 2011.

HONGYU, K.; SANDANIELO, V.L.M. ; OLIVEIRA JUNIOR, G.J. Análise de Componentes Principais: Resumo Teórico, Aplicação e Interpretação. *E&S - Engineering and Science*, v.1, p.83-90, 2016.
KHAYKIN, S. *Neyronnyye seti: polnyy kurs [Neural networks: a comprehensive course].* Williams, Moscow, 2nd ed, p.1104, 2008.

KOHONEN, T. Analysis of a simple self-organizing process. *Biol. Cybern*, v.44 , p.135-140, 1982.

————. *Self-Organization and Associative Memory Process.* Berlin, Springer - Verlag, 1989.

————. Self-Organizing maps: Optimization approaches. *Artificial Neural Networks*, p.891-990, 1991.

RENCHER, A.C. *Methods of Multivariate Analysis.* A John Wiley & Sons, INC. publication, 2ed, p.727, 2002.

SILVA, I.N.; SPATTI D.; FLAUZINO, R. *Redes neurais artificiais para engenharia e ciências aplicadas: curso prático.* Artliber Editora Ltda, São Paulo, SP, Brasil, 2010.

ZOCKUN, M.H.G.P. *A expansão da soja no Brasil: alguns aspectos de produção.* São Paulo: Dissertação de Mestrado em Economia, Universidade de São Paulo, 1978.