Universidade de Brasília
Departamento de Estatística

Análise Multivariada Aplicada a Projeto de Desenvolvimento da Produção Agroecológica no DF

Johnata Alves Moura da Silva
Matrícula: 180020340

Projeto apresentado para o Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Brasília
2023

Johnata Alves Moura da Silva
Matrícula: 180020340

Análise Multivariada Aplicada a Projeto de Desenvolvimento da Produção Agroecológica no DF

Orientador: Prof. Gladston Luiz da Silva

Projeto apresentado para o Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Brasília
2023

Resumo

Neste trabalho, foram estudados os dados dos produtores orgânicos do Distrito Federal, com o objetivo de apresentar análises que possam ser utilizados por um Laboratório de Inovação focado em desenvolver pesquisas para o setor de agricultura orgânica. Aplicou-se Análises Multivariadas de Cluster e de Correspondência em dados coletados de 118 produtores, a fim de verificar a formação de agrupamentos e também traçar um perfil, considerando como variável de classificação o porte dos produtores.

Utilizando método de agrupamento de Ward, foi possível identificar 4 grupos que se diferenciaram na quantidade de produtos utilizados tanto para nutrição vegetal quanto para manejo fitossanitário. A Análise de Correspondência possibilitou verificar a associação entre o porte dos produtores com os sistemas de produção e de manejo.

Palavras-chave: produção orgânica, agroecologia, análise multivariada, análise de correspondência, análise de agrupamentos.

Abstract

This work examined data from organic producers in the Federal District with the aim of presenting analyses that could be used by an Innovation Laboratory focused on developing research for the organic agriculture sector. Multivariate Cluster and Correspondence Analyses were applied to data collected from 118 producers to identify the formation of clusters and to outline a profile, considering the size of the producers as a classification variable.

By employing the Ward clustering method, it was possible to identify 4 groups that presented differences in the quantity of products used for both plant nutrition and phytosanitary management. Correspondence Analysis allowed examination of the association between the size of producers and systems of production and management.

Palavras-chave: organic production, agroecology, multivariate analysis, correspondence analysis, cluster analysis.

Lista de Figuras

1 Mapa do DF com agrupamentos por uso do solo 9
2 Fases do CRISP-DM. 11
3 Área total, área produzida e área de produção orgânica (ha) 24
4 Área total, área produzida e área de produção orgânica (ha) por porte do produtor 24
5 Proporção de área orgânica em hectares por porte do produtor 25
6 Gráfico de Pareto dos produtos de nutrição vegetal. 28
7 Gráfico de Pareto dos produtos de manejo fitossanitário. 28
8 Dendograma pelo método de Ward. 29
9 Áreas de produção e produção orgânica por conglomerado 30
10 Produtos para nutrição vegetal e manejo fitossanitário por conglomerado . 30
11 Projeção em R^2 do Porte, Sistema de Produção e Manejo do Solo 31

Lista de Tabelas

1 Tabela de contingência com i linhas e j colunas. 19
2 Matriz de correspondência das frequências relativas. 20
3 Sistemas de produção por porte 25
4 Manejo do solo por porte 26
5 Manejo de plantas invasoras por porte 26
6 Frequência das categorias de produtos utilizados na nutrição vegetal 27
7 Frequência das categorias de produtos utilizados no manejo fitossanitário. . 27
8 Índices para definição do número de clusters. 29

Glossário

Azadiractina trata-se de um composto químico extraído de sementes de *neen.*

Bokashi adubo constituído de uma mistura de resíduos orgânicos diversos como farinha de osso, farelos de cereais e de oleaginosas que passam por fermentação anaeróbica.

Calda Bordalesa mistura de água, sulfato de cobre e cal. É um insumo utilizado em hortas e pomares orgânicos, devido a sua eficiência, principalmente em controlar várias doenças causadas por fungos (míldio, ferrugem, requeima, pinta preta, cercosporiose, antracnose, manchas foliares, podridões, entre outras) em diversas culturas, tendo efeito secundário contra bacterioses.

Caldas caseiras outras misturas líquidas de diferentes composições químicas mas que, no geral, são produzidas na própia propriedade do produtor orgânico, entre os exemplos mais comuns estão: Calda de pimenta, calda de alho, etc.

Composto produtos que são misturas de compostagem orgânica. Se encaixam nessa categoria resíduos de galinheiro (cama de frango), esterco, restos vegetais etc.

Condicionador/corretivo de solo são produtos que promovem a melhoria das propriedades físicas, físico-químicas ou da atividade biológica do solo.

Controle biológico são as forma de controle de praga que consistem em utilizar organismos vivos que atuam como predadores das pragas. São exemplos de controle biológico: bactérias, vespas, insetos predadores dentre .

Esterco são os produtos provenientes apenas de dejetos de animais, sem misturas.

Fertilizante mineral produto de natureza fundamentalmente mineral, natural ou sintético, obtido por processo físico, químico ou físico-químico, fornecedor de um ou mais nutrientes de plantas.

Reminearlizador são uma categoria de agrominerais silicáticos com diversos grupos, podendo ser ricos em cálcio e magnésio (basálticos), cálcio (calcissilicáticos), magnésio (ultramáficos), potássio (alcalinos) ou cálcio, magnésio e potássio (ultramáficos alcalinos).

Sumário

Capítulo 1 8
1.1 Introdução 8
1.2 Objetivos Geral e Específicos 10
1.3 Metodologia 11
Capítulo 2 - Revisão Bibliográfica 13
2.1 Análise de Conglomerado 13
2.2 Análise de Correspondência 19
2.3 Trabalhos Relacionados 22
Capítulo 3 - Estudo de Caso 23
3.1 Conjunto de Dados 23
3.2 Análises Descritivas 23
3.3 Análise de Conglomerados 29
3.4 Análise Correspondência 31
Conclusão 32

Capítulo 1

1.1 Introdução

Historicamente, a agricultura é uma das práticas intrínsecas do desenvolvimento humano sendo uma das principais formas de sua subsistência. O Brasil é um dos maiores produtores de *commodities*, como soja, cana-de-açúcar, milho e café. Sendo boa parte da produção proveniente da agricultura convencional, de grande porte.

Nesse contexto, a evolução dos meios de produção se mostra necessária para que o crescimento seja sustentável e eficaz. Logo, a produção agroecológica mostra-se uma prática importante para o avanço do setor agrícola. De acordo com a Lei Nº 10.832, de 23 de Dezembro de 2003 que caracteriza a produção orgânica (BRASIL, 2003):

> Art. 1º Considera-se sistema orgânico de produção agropecuária todo aquele em que se adotam técnicas específicas, mediante a otimização do uso dos recursos naturais e socioeconômicos disponíveis e o respeito à integridade cultural das comunidades rurais, tendo por objetivo a sustentabilidade econômica e ecológica, a maximização dos benefícios sociais, a minimização da dependência de energia não-renovável, empregando, sempre que possível, métodos culturais, biológicos e mecânicos, em contraposição ao uso de materiais sintéticos, a eliminação do uso de organismos geneticamente modificados e radiações ionizantes, em qualquer fase do processo de produção, processamento, armazenamento, distribuição e comercialização, e a proteção do meio ambiente.

A produção agrícola também está presente no Distrito Federal (DF). Segundo a EMATER-DF, a área do DF utilizada para agricultura equivale a aproximadamente 4.000 km^2 (quatro mil metros quadrados), quase 70% da área total. As regiões administrativas com maior presença dessa atividade rural são Brazlândia, Planaltina, Gama, São Sebastião e Ceilândia, que até 2015 respondiam por mais de 66% da população rural. São 5.246 estabelecimentos agropecuários e apenas 257 produtores orgânicos (EMATER-DF, 2023; CODEPLAN, 2015).

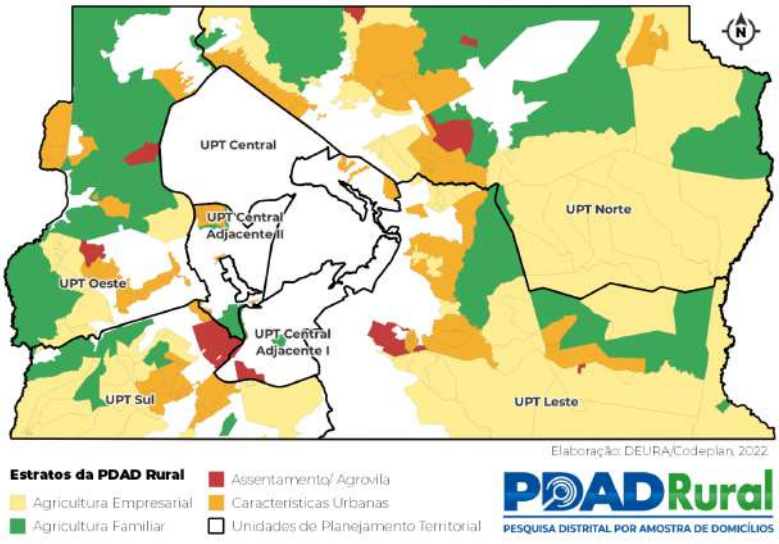

Figura 1: Mapa do DF com agrupamentos por uso do solo

Fonte: PDAD Rural 2022

Com a Figura 1 é possível observar, em verde e amarelo, a área destinada a agricultura, tanto empresarial quanto familiar.

Assim, embora não seja destaque, a produção orgânica tem grande relevância no contexto sustentável da produção de alimentos e, dado que a agricultura convencional evolui de forma insustentável, apresenta-se com uma contraproposta ao sistema de produção predatório. Com isso, surge a necessidade de implementação de novas tecnologias que permitam uma agricultura orgânica mais produtiva, eficiente, sustentável e resiliente (MARCHETTI et al., 2023).

Neste contexto, foi aprovado na FAPDF projeto apresentado pelo Departamento de Engenharia de Produção da Faculdade de Tecnologia da UnB, que resultou na criação do Laboratório de Inovação para estabelecer estratégias e promover iniciativas para desenvolver soluções e novos negócios associados à produção orgânica e de base agroecológica no DF. Tal laboratório é constituído por docentes e discentes da Faculdade de Agronomia e Medicina Veterinária (FAV), Centro de Desenvolvimento Sustentável (CDS) e Faculdade de Tecnologia (FT).

Dessa maneira, este Trabalho de Conclusão de Curso tem como finalidade analisar os dados obtidos através do Laboratório de Inovação referentes às informações dos produtores orgânicos do DF, a fim de contribuir com informações que sejam úteis no desenvolvimento da produção agroecológica.

1.2 Objetivos Geral e Específicos

O objetivo geral deste trabalho é realizar Análises Multivariadas em dados decorrentes do levantado realizado pelo Laboratório de Inovação junto aos produtores que integram a cadeia de produção de orgânicos e agroflorestais do DF, com o objetivo de subsidiar o Laboratório de Inovação com informações que contribuam com o desenvolvimento de iniciativas que reforcem a produção agroecológica do Distrito Federal.

Para alcançar o objetivo geral, este projeto tem os seguintes objetivos específicos:

- Realizar Análise Exploratória para identificar variáveis com poder de discriminação.
- Realizar Análises de Correspondência para obter uma representação multivariada de interdependência para dados não-métricos.
- Realizar Análises de Agrupamentos para classificar indivíduos ou objetos em um número de grupos mutuamente excludentes, com base nas similaridades entre as entidades.

1.3 Metodologia

Para a estrutura da análise dos dados e desenvolvimento de possíveis modelos, foi utilizada a metodologia CRISP-DM (Acrônimo de *CRoss-Industry Standard Process for Data Mining*) (PETE et al., 2000). Esse método é composto por 6 fases como ilustrado na Figura 2.

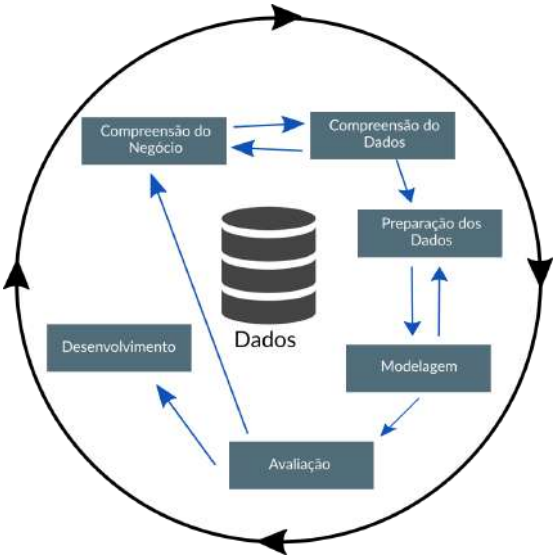

Figura 2: Fases do CRISP-DM.

Fonte: (PETE et al., 2000) com adaptações

As etapas são desenvolvidas da seguinte maneira:

1. Compreensão do negócio: uma série de objetivos que visam estruturar a necessidade, os objetivos e a finalidade da mineração de dados;
2. Entendimento dos dados: coleta inicial dos dados, onde será feita a primeira análise exploratória com o objetivo de determinar sua qualidade;
3. Preparação dos dados: os dados selecionados na fase anterior serão tratados para a realização dos passos seguintes;
4. Modelagem: seleção e treinamento de modelos que possam produzir resultados úteis para a solução do problema;
5. Avaliação: os modelos selecionados anteriormente são comparados entre si a partir da precisão de cada um;
6. Implementação: fase final, onde o melhor modelo é aplicado e monitorado a fim de observar a adaptação do modelo.

No Capítulo 2 é apresentado o referencial teórico, além de alguns trabalhos relacionados na área de produção orgânica com o intuito de inserir o presente trabalho nesse de contexto, além de explicitar a importância da pesquisa para o avanço do setor orgânico.

Capítulo 2 - Revisão Bibliográfica

2.1 Análise de Conglomerados

A Análise de Conglomerado é uma técnica que consiste em aplicar algum algoritmo de divisão a fim de criar grupos, de forma que os objetos dentro de um mesmo grupo se assemelham entre si, enquanto que objetos de grupos diferentes são distintos (GAN et al., 2007). Também uma forma de análise exploratória, ao analisarmos diferentes estruturas do banco de dados é possível identificar valores discrepantes e possíveis relações entre as variáveis (JOHNSON; WICHERN, 2007).

Alguns altores também definem as técnicas de Conglomerados como aprendizado não supervisionado, uma vez que procura-se uma nova forma de estruturar os dados. Um dos desafios desse tipo de analise é a falta de comprovação, validação, dos resultados (JAMES et al., 2013; IZENMAN, 2009).

A técnica consiste em, a partir de medidas de similaridade entres os objetos, aplicar um algoritmo que minimize as distâncias dentro dos grupos. São vários os métodos de clusterização, mas que podem ser divididos entre os hierárquicos e não hierárquicos.

As medidas de similaridade ou dissimilaridade são a base da classificação dos clusters e medem as distâncias entre as observações duas a duas. Uma métrica comum é a distância Euclidiana entre dois objetos, definida como:

$$d(x, y) = \sqrt{\sum_{j=1}^{p} (x_j - y_j)^2} \tag{2.1.1}$$

onde $\mathbf{x}$ e $\mathbf{y}$ são vetores linha de tamanho p. Outra medida conhecida é a de Minkowski, definida por:

$$d_m(x, y) = \left[\sum_{j=1}^{p} |x_j - y_j|^m \right]^{\frac{1}{m}} \tag{2.1.2}$$

Sendo a distância Euclidiana um caso especial quando m=2.

Os métodos hierárquicos podem ainda ser divididos em aglomerativos e divisivos. Os aglomerativos consistem em agrupar as observações a partir de n grupos, um para cada observação, então agrupa-se os objetos mais similares e os grupos mais similares até chegarmos em um único grupo. Os métodos divisivos têm uma ideia contrária, isto é, partindo de um conglomerado com todos os objetos faz-se dois grupos

de forma que os objetos em um sejam os mais distantes do outro repetindo-se até chegar em n grupos. Para ambos os casos pode-se construir um diagrama de árvore (dendograma) que ilustra a distância entre cada objetos e os grupos formados (JOHNSON; WICHERN, 2007).

Dos métodos aglomerativos temos alguns como: Ligação Única (*Single Linkage*), Ligação Completa (*Complete Linkage*), Ligação Média (*Average Linkage*) e o Método de Ward que são citados por Johnson e Wichern (2007), Mardia et al. (1979), Gan et al. (2007) dentre outros autores.

Ligação Única

No primeiro passo são calculadas as distância entre os objetos e mescla-se os 2 objetos com a menor distância, ou seja, considerando U e V dois objetos com a menor distância em relação ao demais, o conglomerado (UV) denota o grupo formado por U e V. Em seguida, calcula-se as distância de UV e qualquer outro grupo W:

$$d_{(UV)W} = min\{d_{UW}, d_{VW}\} \tag{2.1.3}$$

Por fim, os resultados podem ser visualizados em um dendograma e a escolha da quantidade de conglomerados pode ser feita com base nas distâncias entre cada grupo gerado.

Ligação Completa

Semelhante ao anterior, porém considerando a maior distância entre dois elementos de conglomerados distintos, garantindo que todos os objetos de um grupo estão dentro de uma distância máxima um do outro. Assim, as distâncias entre (UV) e outro conglomerado W são dadas por:

$$d_{(UV)W} = max\{d_{UW}, d_{VW}\} \tag{2.1.4}$$

Ligação Média

Para este caso, após o primeiro agrupamento a partir das distâncias mínimas, as distâncias entre (UV) e outro conglomerado W são dadas por:

$$d_{(UV)W} = \frac{\sum_i \sum_k d_{ik}}{N_{(UV)} N_W} \tag{2.1.5}$$

onde d_{ik} é a distância entre um objeto i do grupo (UV) e o objeto k do grupo W, e $N_{(UV)}$ N_W são os números de itens em cada conglomerado.

Método de Ward

Se diferenciando dos citados anteriormente, o método de Ward procura minimizar a perda de informação, isto é, o aumento do erro da soma de quadrados é mínimo. Assim, para um certo conglomerado k, seja ESQ_k o desvio do erro da soma de quadrados de cada elemento em relação à média do grupo. Assim, considerando K agrupamentos, o ESQ é definido como a soma dos ESQ_k. Assim, em cada passo do método, a união de cada par de grupos é considerada e os dois conglomerados em que o ESQ seja mínimo são mesclados. Sendo um método aglomerativo, inicialmente temos n grupos, de forma que o $ESQ_k = 0$ assim como o ESQ. Quando todos os grupos são combinados, o ESQ é dado por:

$$ESS = \sum_{j=1}^{N} (x_j - \bar{x})'(x_j - \bar{x}) \tag{2.1.6}$$

onde x_j é a medida multivariada da j-ézima observação e $\bar{x}$ é a média de todas as observações. Por fim, os resultados podem ser observados também por um dendograma, mas o eixo vertical indica o ESQ de cada união.

Gan et al. (2007) apresenta ainda alguns dos métodos divisivos como: DIANA, DISMEA.

DIANA (*Divisive Analisys*)

Este é aplicado a partir de uma série de divisões sucessivas. Em cada passo, o conglomerado com maior distância entre dois objetos é dividido até que o passo $n - 1$ em que os grupos são unitários.

Considerando C o grupo a ser dividido, A e B os grupos resultantes da divisão de C. O algoritmo encontra A e B movendo ponto entre eles de forma iterativa. Na primeira etapa, o ponto y_i será movido de A para B se maximizar a função:

$$D(x, A\backslash\{x\}) = \frac{1}{|A| - 1} \sum_{y \in A, y \neq x} d(x, y) \tag{2.1.7}$$

onde, $A\backslash\{x\}$ denota o grupo A menos a observação x, $|A|$ é o número de item no grupo A e $d(\cdot, \cdot)$ pode ser qualquer medida de distância apropriada para os dados. Nos passos seguintes, procura-se outros elementos de A que deveriam ser alocados em B

seguindo a função:

$$D(x, A\backslash\{x\}) - D(x, B) \tag{2.1.8}$$

Assim, se um ponto, seja, y_2 maximiza a equação (2.1.8) então y_2 será movido de A para B. Se o valor for negativo ou igual a 0, a divisão está completa.

DISMEA

É um algoritmo divisivo que usa o algoritmo *k-means* para dividir um grupo em dois. Em cada passo, o grupo com maior soma de quadrado das distâncias (SQD) é dividido, fazendo isso até chegar em n grupos unitários. A primeira etapa é encontra um bipartição inicial C_1 e C_2 que minimize a função:

$$F(C_1, C_2) = \sum_{i=1}^{2} \sum_{x \in C_i} \parallel x - \mu(C_i) \parallel^2 \tag{2.1.9}$$

onde $\mu(C_i)$ é o centroide (média das distâncias) do conglomerado C_i. Nos passos seguinte, escolhe-se o grupo com maior SQD para ser dividido, onde o SQD para um grupo C é dado por:

$$E(C) = \sum_{x \in C} \parallel x - \mu(C) \parallel^2 \tag{2.1.10}$$

K-means

Dos métodos não hierárquicos, o mais citado na literatura seria, talvez, o *K-means*. O método realiza diversas iterações, trocando os objetos entre os grupos com médias (centroides) mais próximas até que a função de erro não tenha alteração significativa. Nesse caso, o número de grupos é predeterminado.Assim, seja $C_1, C_2, \cdots, C_k$ k conglomerados, a função de erro é dada por:

$$E = \sum_{i=1}^{n} \sum_{x \in C_i} d(x, \mu(C_i)) \tag{2.1.11}$$

onde $d(x, \mu(C_i))$ denota a distância entre x e $\mu(C_i)$, podendo ser qualquer métrica de distância, mas tipicamente escolhe-se a distância euclidiana. O *k-means* tem ainda algumas variações no algoritmo que, em sua maioria, buscam otimizar ainda mais o método (GAN et al., 2007).

Por fim, ainda pode-se usar alguma métrica para encontra o número ótimo de agrupamentos. Charrad et al. (2014) apresenta diversos índices, juntamente com a implementação no R por meio do pacote 'NbClust'. Esses índices podem servir como um indicativo, mas a escolha também pode ser feita seguindo outros critérios ou a natureza do banco de dados. Para o cálculo dos índices, considere as seguintes notações:

- n = número de observações;
- p = número de variáveis;
- q = número de clusters;
- x =$\{x_{ij}\}$, i = 1,2, ..., n, j = 1,2 ,...,p
  = matriz n x p de p variáveis medidas e n observações independentes;
- $\bar{X}$ = matriz q x p da médias dos clusters;
- x = centroide da matriz de dados X;
- n_k = número de objetos no cluster C_k;
- c_k = centroide do cluster Ck;
- x_i = vetor p-dimensional de características do i-ésimo objeto no grupo C_k;
- $W_q = \sum_{k=1}^{q} \sum_{i \epsilon C_k} (x_i - c_k)(x_i - c_k)^T$ é a matriz de dispersão dentre grupos para dados agrupados em q clusters;
- $B_q = \sum_{k=1}^{q} n_k (c_k - \bar{x})(c_k - \bar{x})^T$ é a matriz de dispersão entre grupos para dados agrupados em q clusters;
- N_t = número de pares de observações no banco de dados: $N_t = \frac{n(n-1)}{2}$;
- N_w = número de pares de observações pertencentes ao mesmo grupo:
  $N_w = \sum_{k=1}^{q} \frac{n_k(n_k-1)}{2}$;
- N_b = número de pares de observações pertencentes à grupos diferentes:
  $N_b = N_t - N_w$;
- S_w = soma das distancias dentro dos clusters: $S_w = \sum_{k=1}^{q} \sum_{i,j \epsilon C_k} d(x_i, x_j)$ com i < j;
- S_b = soma das distancias entre os clusters: $S_b = \sum_{k=1}^{q-1} \sum_{l=k+1}^{q} \sum_{i \epsilon C_k,\ j \epsilon C_l} d(x_i, x_j)$;

E então, alguns indicadores são definidos como:

- **Índice de Calinski e Harabasz** (CH index):

$$CH(q) = \frac{tr(B_q)/(q-1)}{tr(W_q)/(n-q)} \tag{2.1.12}$$

e o valor q que maximiza CH indica o número de agrupamentos (CALINSKI; HARABASZ, 1974);

- **Índice Duda**: Duda e Hart (1973) propõe a razão (2.1.13), sendo a soma de quadrados de dois grupos, seja C_k e C_l, particionados de outro C_m:

$$Duda = \frac{W_k + W_l}{W_m} \tag{2.1.13}$$

Gordon (1999) propõe que o número ótimo de grupos é dado por:

$$Duda \geq 1 - \frac{2}{\pi p} - 3,2\sqrt{\frac{2(1-\frac{8}{\pi^2 p}}{n_m p}} = critValue\_Duda \tag{2.1.14}$$

- **Pseudo** T^2: Duda e Hart (1973) também propõe outro índice, porém sendo indicado para métodos :

$$Pseudot2 = \frac{W_m - W_k - W_l}{\frac{W_k+W_l}{n_k+n_l-2}} \tag{2.1.15}$$

Gordon (1999) especifica que o número ótimo de grupos é menor q onde:

$$Pseudot2 \leq \left(\frac{1 - critValue\_Duda}{critValue\_Duda}\right) \times (n_k + n_l - 2) \tag{2.1.16}$$

- **indice CCC**: o critério de agrupamento cúbico (*Cubic Clustering Criterion*) (CCC), apresentado por Sarle (1983), é dado por:

$$CCC = ln\left[\frac{1 - E(R^2)}{1 - R^2}\right] \frac{\sqrt{np^*/2}}{(0,001 + E(R^2))^{1,2}} \tag{2.1.17}$$

onde

$$R^2 = 1 - \frac{tr(X^T X - \bar{X}^T Z^T Z \bar{X})}{tr(X^T X)} \tag{2.1.18}$$

e o valor máximo do índice é utilizado para obter o número ótimo de agrupamentos.

Além desses critérios, pode-se usar o próprio dendograma (nos casos hierárquicos) para definir o número de grupos, geralmente escolhe-se o ponto de corte onde há um salto no gráfico, isto é, as distâncias entre os grupos são mais altas.

2.2 Análise de Correspondência

A Análise de Correspondência é uma técnica que faz parte da Análise Multivariada e tem aplicabilidade na análise de dados categóricos. Utilizada tanto para a redução dimensional quanto para representação gráfica e também útil na visualização de correlação por meio das distâncias qui-quadrado, de forma que a associação entre as variáveis pode ser retratada como a proximidade entre os pontos dos gráfico (HAIR et al., 2009).

Essa análise toma como ponto de partida uma tabela de contingência, onde calcula-se a diferença entre a frequência esperada e a frequência observada, padroniza-se essa diferença por meio do valor qui-quadrado e a converte para uma medida de associação, como é visto em Hair et al. (2009).

Assim, para se aplicar a decomposição em valores singulares (DVS, ou singular value decomposition - SVD) primeiro é calculada a matriz de correspondência, que consiste em dividir o valor de cada célula pelo total na tabela de contingência.

Tabela 1: Tabela de contingência com i linhas e j colunas.

Linhas	Colunas				Total da linha
	1	2	$\cdots$	j	
1	n_{11}	n_{12}	$\cdots$	n_{1j}	$n_{1.}$
2	n_{21}	n_{22}	$\cdots$	n_{2j}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
i	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$n_{i.}$
Total da coluna	$n_{.1}$	$n_{.2}$	$\cdots$	$n_{.j}$	n

Fonte: Greenacre (2007)

Tabela 2: Matriz de correspondência das frequências relativas.

Linhas	Colunas				Total da linha
	1	2	$\cdots$	j	
1	p_{11}	p_{12}	$\cdots$	p_{1j}	$p_{1.}$
2	p_{21}	p_{22}	$\cdots$	p_{2j}	$p_{2.}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
i	p_{i1}	p_{i2}	$\cdots$	p_{ij}	$p_{i.}$
Total da coluna	$p_{.1}$	$p_{.2}$	$\cdots$	$p.j$	1

Fonte: Greenacre (2007)

A Tabela 1 pode ser denotada por $\mathbf{N}$ e a tabela 2 é denotada por $\mathbf{P} = (p_{ij}) = (n_{ij}/n) = \mathbf{N}/n$ e retrata as frequências relativas observadas. A partir da matriz $\mathbf{P}$ são calculados os perfis das linhas e das colunas:

$$r_i = \sum_{j=1}^{J} p_{ij} = \sum_{j=1}^{J} \frac{n_{ij}}{n}, i = 1, 2, \cdots, I, \ ou \ \underset{(i\times 1)}{\mathbf{r}} = \underset{(i\times j)(j\times 1)}{\mathbf{P}} \ \mathbf{1} \tag{2.2.1}$$

$$c_j = \sum_{i=1}^{I} p_{ij} = \sum_{i=1}^{I} \frac{n_{ij}}{n}, j = 1, 2, \cdots, J, \ ou \ \underset{(j\times 1)}{\mathbf{c}} = \underset{(j\times i)(i\times 1)}{\mathbf{P'}} \ 1 \tag{2.2.2}$$

Também podem ser definidas as frequências esperadas:

$$E_{ij} = \frac{n_{i.} \times n_{.j}}{n_{..}} \tag{2.2.3}$$

E então, é calculada estátistica qui-quadrado:

$$\chi^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{2.2.4}$$

Que tem distribuição qui-quadrado com (1-I)(1-J) sob suposição de independência entre as variáveis.

Johnson e Wichern (2007) também definem as matrizes diagonais dos perfis e aplica o inverso da raiz quadrada para propósitos de padronização:

$$\mathbf{D}_r^{-1/2} = diag\left(\frac{1}{\sqrt{r_1}}, \cdots, \frac{1}{\sqrt{r_i}}\right) \ e \ \ \mathbf{D}_c^{-1/2} = diag\left(\frac{1}{\sqrt{c_1}}, \cdots, \frac{1}{\sqrt{c_j}}\right) \tag{2.2.5}$$

E Greenacre (2007) define um passo a passo de como obter as coordenadas dos perfis de linhas e colunas, onde primeiro calcula-se a matriz S dos resíduos padronizados:

$$S = D_r^{-\frac{1}{2}}(P - rc^T)D_c^{-\frac{1}{2}} \tag{2.2.6}$$

Calcular o SVD de S:

$$S = UD_\alpha V^T \ onde \ \ U^TU = V^TV = I \tag{2.2.7}$$

sendo D_α a diagonal dos valores singulares positivos em ordem decrescente: $\alpha_1 > \alpha_2 \cdots$

Calcular as coordenadas padronizadas das linhas e colunas:

$$\Phi = D_r^{-\frac{1}{2}}U \ e \ \Gamma = D_c^{-\frac{1}{2}}V \tag{2.2.7}$$

E as coordenadas principais das linhas e colunas:

$$\Phi = \Phi D_\alpha \ e \ \Gamma = \Gamma D_\alpha \tag{2.2.8}$$

E por fim, as inercias parciais são dadas por:

$$\lambda_k = \alpha_k^2, k = 1, 2, ..., min\{I - 1, J - 1\} \tag{2.2.9}$$

Que são as quantidades de variância explicada de cada componente.

2.3 Trabalhos Relacionados

Na literatura, de uma maneira geral, são encontrados trabalhos focados em aprimoramento de recursos e\ou processos, como trata a autora em Silva et al. (2014), onde são estudadas metodologias para combate à fungos em plantações.

Artigos como Sabourin et al. (2019) tratam do desenvolvimento de políticas públicas a favor da agroecologia e, nesse caso, o estudo é feito no DF. O autor parte de uma metodologia de entrevistas e análise de documentos e visa encontrar vetores importantes no desenvolvimento da produção orgânica no DF.

Também no contexto de Distrito Federal e RIDE, Muñoz et al. (2022) analisa a sustentabilidade a longo prazo de três unidades familiares, onde foram calculados índices de sustentabilidade. O autor concluiu a pesquisa com índices positivos, porém apontou a necessidade de investimento em solo sob manejo orgânico. Logo, tem-se um reforço para a criação de tecnologias e desenvolvimento de estudos que alavanquem e aprimorem o setor de produção orgânica do DF.

Fenner et al. (2022) também discute o avanço da agroecologia sobre a temática da comparação entre agroecologia e agrotóxicos e sob a ótica de Territórios Saudáveis e Sustentáveis (TSS). A pesquisa é feita por revisão de literatura e análise de documentos pertinentes e também reforça a necessidade de aprofundamento nas pesquisas para melhorar a relação entre o TSS e agroecologia.

Assim, diante do exposto, vê-se a necessidade de mais trabalhos focados em desenvolver melhorias e prestar suporte ao setor agroecológico. Embora o presente trabalho (assim como o laboratório de pesquisa) tenha enfoque apenas na região do DF, as análises podem ser replicadas para outras regiões ou, até mesmo, aplicadas à nível nacional. As escolhas das técnicas discutidas a seguir, tem como objetivo apresentar uma análise preliminar, mostrando as relações entre o tamanho dos produtores, os sistemas de produção utilizados, as técnicas de manejo de solo e de plantas invasoras e os produtos (ou maneiras) que são utilizados para nutrição e manejo fitossanitário (prevenção de pragas e doenças).

Capítulo 3 - Estudo de Caso

A seguir, serão apresentadas as análises desenvolvidas acerca dos dados obtidos pelo Laboratório de Inovação através dos planos de manejo dos produtores vinculados à OPAC cerrado. Nas etapas da metodologia CRISP-DM, esta seria a etapa de entendimento dos dados, preparação dos dados, modelagem e avaliação. A etapa de compreensão está presente no Capítulo 1 desse trabalho.

3.1 Conjunto de Dados

Iniciando a etapa de entendimento dos dados, o banco contém informações de 118 produtores, que possuem uma área de aproximadamente 1971 ha (hectares), dos quais 724 ha são destinadas à produção orgânica. Ao todo são 7 produtores de grande porte, 11 de médio porte, 87 de pequeno porte e 13 minifúndios.

As variáveis obtidas são relacionadas ao tipo de produção, manejo do solo, tipo de nutrição vegetal, manejo fitossanitário, manejo de plantas invasora e o processamento da produção, totalizando 39 variáveis.

Inicialmente, foram analisados dados relativos aos produtos de manejo fitossanitário e nutrição vegetal a fim de analisar fatores em comum entre os produtores. Para isso, foi elaborada uma categorização dos produtos e então aplicada transformação para variáveis *dummies*. Assim, foi possível identificar produtos mais usados e que mais representavam os produtores.

3.2 Análises Descritivas

Ainda na etapa de entendimento dos dados, pensando no contexto dos dados em estudo, a análise do perfil dos produtores é crucial para o entendimento do contexto do estudo. Das variáveis disponíveis, quatro são relativas à área: a área total da propriedade, a área utilizada para produção, área utilizada para a produção orgânica e porte do produtor (essa última sendo relacionada à área total).

A Figura 3 apresenta o tamanho das propriedades, a parte destinada à produção em geral e as terras destinadas apenas para a produção orgânica de fato. Observa-se que a tendencia é que as barras do gráfico fiquem maiores para áreas menores, isto é, as porções de terras destinas para a produção geral e orgânica são menores que o tamanho total da propriedade. De uma maneira geral, os tamanhos das propriedades se concentram em menos 50 ha. Nas áreas totais, observam-se três barras ao extremo, acima de 200 ha, que nos gráficos de área produzida 'somem', indicando que

esses produtores utilizam menos da metade do tamanho total da propriedade.

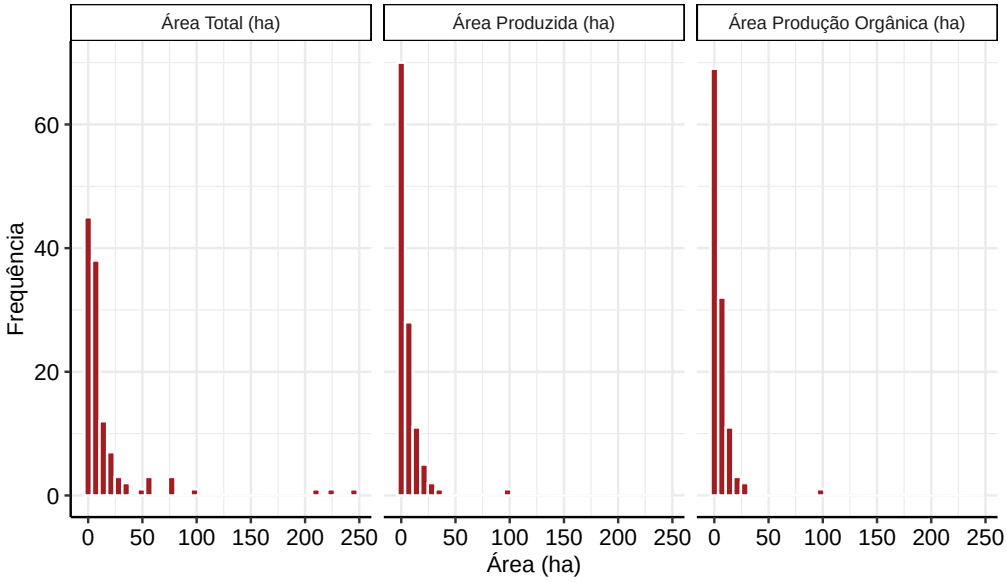

Figura 3: Área total, área produzida e área de produção orgânica (ha)

Com a Figura 4 é possível observar os tamanhos das propriedade dentro dos grupos de porte. Essa categorização dos produtores já fazia parte do banco de dados e é feita com base na área total. Assim, é possível observar que cada porte é bem dividido nos *box plots* da área total. Ainda, os valores discrepantes observados na Figura 3 fazem com que a média da área total do grande porte fique distante da mediana. Nas áreas produzidas observa-se que médio e frande porte ficam logo abaixo dos 50 ha, com um produtor destoante com 100 ha de área produzida.

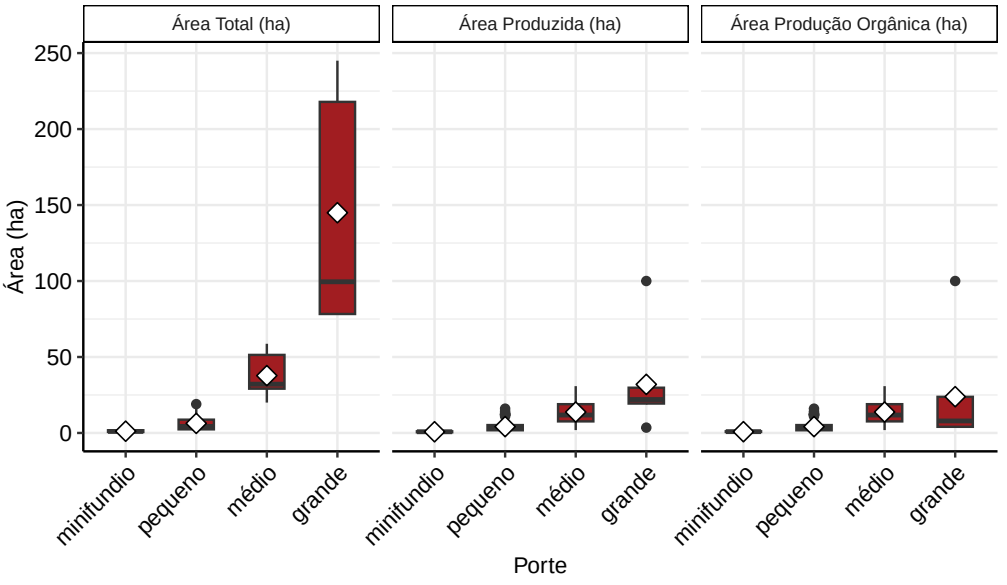

Figura 4: Área total, área produzida e área de produção orgânica (ha) por porte do produtor

A Figura 5 apresenta a proporção destinada a produção orgânica em relação ao tamanho total da propriedade para cada tipo de porte. E assim, pode-se notar que a grande maioria dos produtores de médio e grande porte utiliza menos de 50% da área total para a produção de orgânicos.

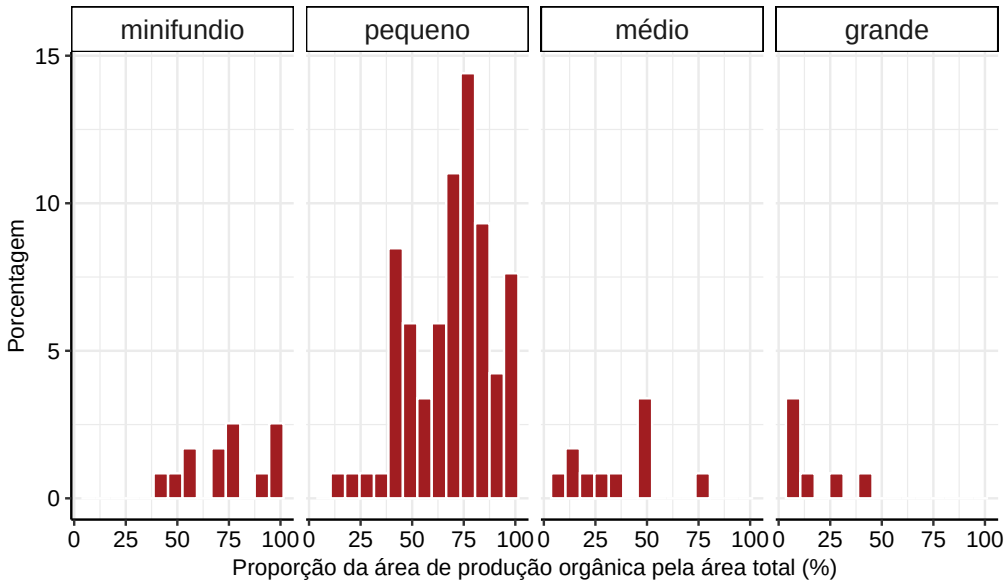

Figura 5: Proporção de área orgânica em hectares por porte do produtor

Também são analisados os tipos de produção, manejo de solo, manejo de plantas invasoras, produtos utilizados para nutrição vegetal e produtos utilizados para manejo fitossanitário. Os três primeiros são apresentados no banco de dados por variáveis binárias, enquanto que os produtos passaram por uma categorização devido à grande diversidade de produtos.

No que diz respeito aos sistemas de produção, a Tabela 3 apresenta os sistemas de produção por porte dos produtores. É importante ressaltar que o mesmo produtor pode utilizar mais de um sistema e a categoria "Grãos" também engloba outros tipos de culturas anuais.

Tabela 3: Sistemas de produção por porte

Porte	**Fruticultura**		**Grãos**		**Hortaliça**	
	n	(%)	n	(%)	n	(%)
Grande	5	(71)	7	(100)	4	(57)
Médio	11	(100)	10	(99)	11	(100)
Pequeno	86	(99)	76	(87)	86	(99)
Minifúndio	12	(99)	9	(69)	12	(99)
Total	114	(97)	102	(96)	113	(96)

A Tabela 4 apresenta os tipos de manejo do solo utilizados pelos produtores. Nota-se que apenas 41% dos produtores utilizam curvas de nível enquanto que as demais categorias são mais frequentes.

Tabela 4: Manejo do solo por porte

Porte	**Adubação Verde**		**Curva de Nível**		**Manutenção de restos vegetais**		**Rotação de culturas**	
	n	(%)	n	(%)	n	(%)	n	(%)
Grande	7	(100)	3	(43)	7	(100)	4	(57)
Médio	9	(82)	4	(36)	9	(82)	11	(100)
Pequeno	80	(92)	38	(44)	83	(95)	80	(92)
Minifúndio	7	(54)	4	(31)	13	(100)	12	(99)
Total	103	(87)	49	(42)	112	(95)	107	(91)

A Tabela 5 exibe a frequência de cada meio de manejo de plantas invasoras. De uma maneira geral, os produtores retiram as plantas invasoras manualmente. As práticas de *Muching* (cobrir o solo não plantado com plástico) e redução no banco de sementes são menos utilizadas.

Tabela 5: Manejo de plantas invasoras por porte

Porte	**Controle Mecânico**		**Manual**		**Muching**		**Redução de Sementes**	
	n	(%)	n	(%)	n	(%)	n	(%)
Grande	7	(100)	7	(100)	2	(29)	3	(43)
Médio	9	(82)	11	(100)	5	(45)	1	(9)
Pequeno	54	(62)	85	(98)	38	(44)	40	(46)
Minifúndio	6	(46)	13	(100)	5	(38)	7	(54)
Total	76	(64)	116	(98)	50	(42)	51	(43)

As Tabelas 6 e 7 apresentam as as frequência dos produtos que são utilizados para nutrição vegetal e manejo fitossanitário. 5 produtores não apresentaram informações de produtos fitossanitários, 1 produtor não apresentou produtos de nutrição e 1 produtor não apresentou produtos para ambos.

Essa análise é um pouco imprecisa, principalmente devido à categorização dos produtos que foi feita de forma manual e em conjunto com pessoas com tal conhecimento técnico. Além disso, alguns produtos são utilizados tanto para nutrição quanto para manejo de pragas e doenças, isto é, alguns produtos de nutrição também tinham efeito no controle fitossanitário mas como uma consequência.

Tabela 6: Frequência das categorias de produtos utilizados na nutrição vegetal

Produto	**Frequência**	**Frequência Relativa**	**Frequência Acumulada**
Remineralizador	88	11,1%	11,1%
Composto	79	9,9%	21,0%
Bokashi	71	8,9%	30,0%
Condicionador/ corretivo de solo	68	8,6%	38,5%
Esterco	65	8,2%	46,7%
Fertilizante mineral	61	7,7%	54,4%
Demais categorias	362	45,6%	

NOTA: A descrição dos grupos encontra-se no glossário

Tabela 7: Frequência das categorias de produtos utilizados no manejo fitossanitário.

Produto	**Frequência**	**Frequência Relativa**	**Frequência Acumulada**
Controle biológico	82	13,7%	13,7%
Calda bordalesa	66	11,0%	24,7%
Azadiractina	54	9,0%	33,8%
Caldas caseiras	47	7,9%	41,6%
Enxofre	41	6,9%	48,5%
Fertilizante	35	5,9%	54,3%
Demais categorias	213	46%	

NOTA: A descrição dos grupos encontra-se no glossário

A Figura 6 realça, ainda, 70% dos produtores utilizam 7 categorias de produtos, enquanto 80% utilizam 12 categorias diferentes.

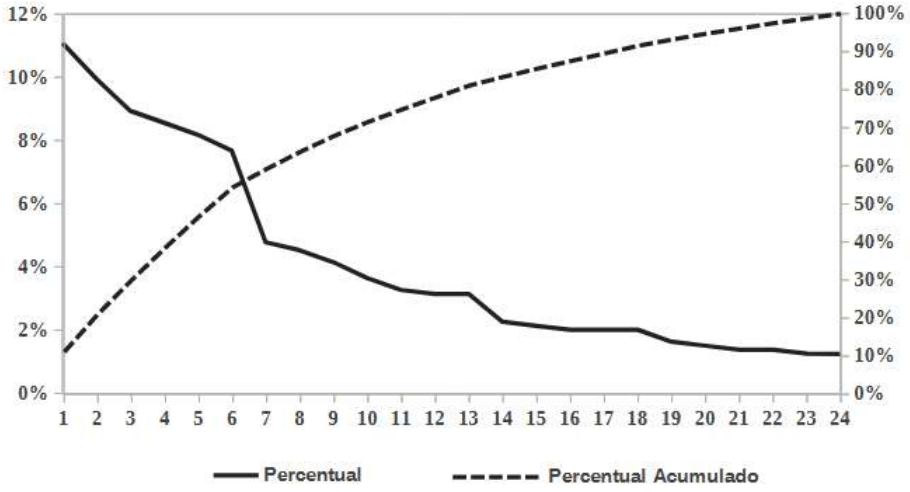

Figura 6: Gráfico de Pareto dos produtos de nutrição vegetal.

Para os produtos de manejo fitossanitário, a Figura 7 expõe que 8 categorias são usadas por 70% das propriedade, ao mesmo tempo que 80% dos produtores utilizam 11 categorias.

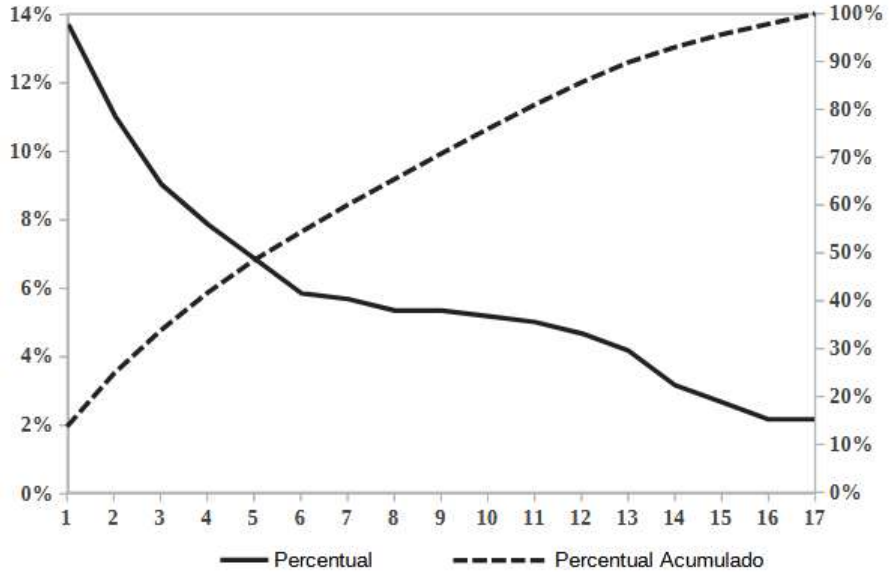

Figura 7: Gráfico de Pareto dos produtos de manejo fitossanitário.

3.3 Análise de Conglomerados

Diante do exposto na seção anterior, buscou-se então reagrupar os produtores considerando os tamanhos de áreas destinadas à produção e produção orgânica, bem como a quantidade de produtos utilizados para nutrição vegetal e manejo fitossanitário.

A Tabela 8 apresenta os números de conglomerados por índice. Nota-se que existe uma diferença na quantidade de grupos indicados e assim, optou-se por escolher o número de conglomerados através do Dendrograma. Vale salientar que foram considerados 111 produtores, uma vez que haviam produtores sem dados a respeito dos produtos utilizados.

Tabela 8: Índices para definição do número de clusters.

Índice	**Nº de cluster**	**Valor do Índice**
Duda	2	0,84
PseudoT2	2	13,32
CH	6	83,96
CCC	9	1,96

Assim, considerando os valores escalonados das variáveis selecionadas, o método de Ward e as distâncias euclidianas. A Figura 8 apresenta o dendograma obtido. Nota-se que uma discrepância na quantidade de produtores por grupo e, junto à isso, um grupo é composto por apenas 1 produtor que tem 100 ha de produção.

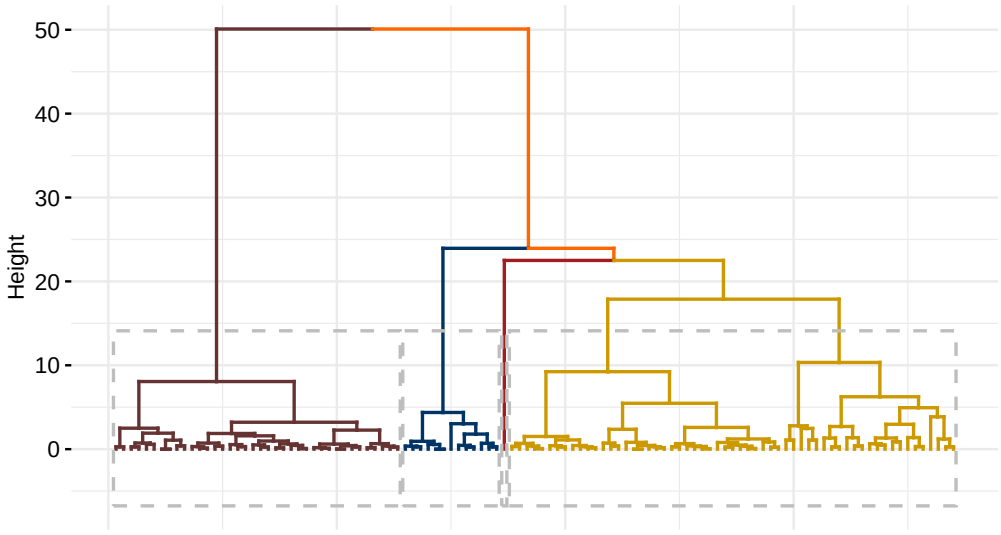

Figura 8: Dendograma pelo método de Ward.

As Figuras 9 e 10 exibem as distribuições das áreas de produção, produção orgânica e quantidades de produtos de nutrição vegetal e fitossanitário por grupo. É observado que os tamanhos das áreas ainda são próximos entre os conglomerados, mas o grupo 3 têm médias menores e, desconsiderando o grupo 4, o grupo 2 têm médias maiores. Já para a quantidade de produtos, esses são mais diferenciados entre os conglomerados. Os produtores do conglomerado 1 tem menor variedade de produtos de ambos os tipos, enquanto que o conglomerado 3 tem maior variedade (salientando que são considerados produtos em categorias).

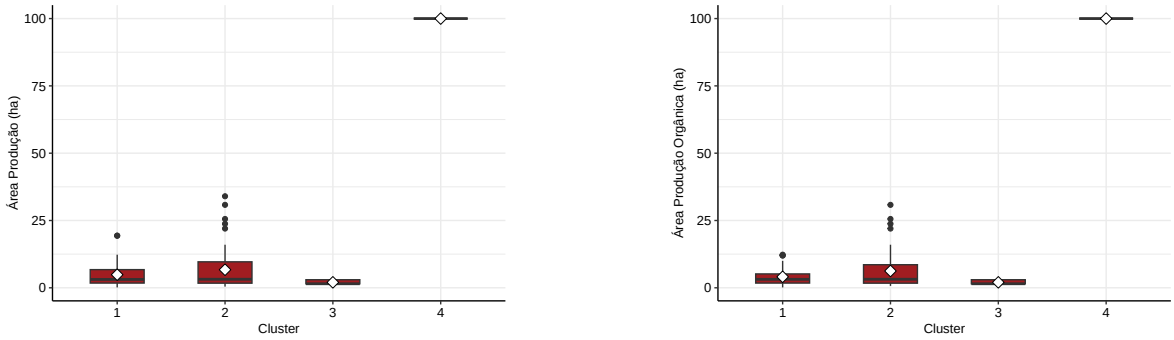

Figura 9: Áreas de produção e produção orgânica por conglomerado

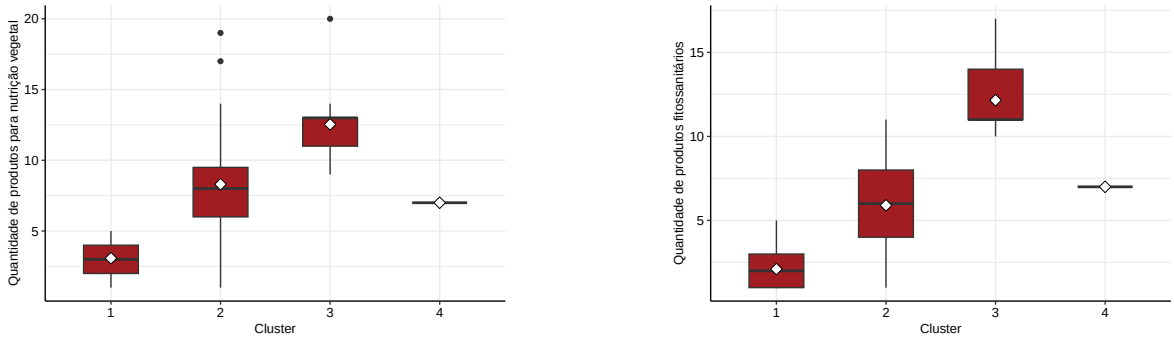

Figura 10: Produtos para nutrição vegetal e manejo fitossanitário por conglomerado

Assim, embora os grupos não sejam muito heterogêneos em relação aos tamanhos das produções, é possível separar clusters por diversidade de produtos que são utilizados tanto para nutrição vegetal quanto para manejo fitossanitário.

3.4 Análise Correspondência

A Figura 11 apresenta o Gráfico de Correspondência entre as variáveis citadas. Inicialmente, observa-se que a Dimensão 1 explica quase 75% da variabilidade dos dados indicando que as distâncias horizontais, ou em relação ao eixo da Dimensão 1, são mais representativas. Também, observa-se que os produtores de pequeno porte são bem próximos da origem, indicando que esses não são muito representativos. De fato, os produtores de pequeno porte apresentam poucas diferenças nas frequências das categorias de sistemas e manejo. Minifúndio e Grande porte são os mais distantes da origem, sendo assim as categorias de porte mais destoantes das demais e entre si.

Além disso, observando relações entre porte e sistemas de produção, nota-se os produtores de Grande porte mais próximos com a produção de Grãos. Os produtores de Médio porte são mais relacionados ao sistema de Hortaliça e Fruticultura. Já os produtores de Pequeno porte, por representarem a maior parte da amostra, têm frequências maiores dos 3 tipos de produção e consequentemente também apresenta maior proximidade dos 3. O Minifúndio apresenta maior proximidade com sistemas de Fruticultura e Hortaliça.

Considerando as variáveis de Manejo de solo, pode-se observar que os produtores de Grande porte, embora sejam distantes de todas as categorias, aparentam uma distância menor da adubação Verde. O Médio tem maior relação com o Manejo por Rotação de culturas. Os produtores de Pequeno porte são mais próximos de curvas de nível do que os demais, também utilizam Adubação Verde e Rotação de Culturas. Por fim, os produtores do Minifúndio são os mais próximos do manejo por Manutenção de Restos Vegetais.

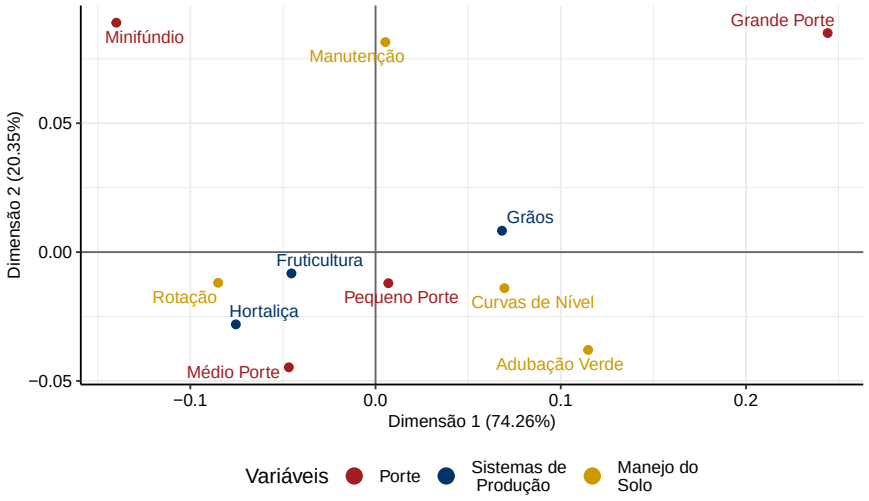

Figura 11: Projeção em R^2 do Porte, Sistema de Produção e Manejo do Solo

Conclusão

Este trabalho teve como principal objetivo estudar os dados relativos aos produtores orgânicos do DF. Utilizando a base de dados coletados pelo Laboratório de Inovação, composta por 118 produtores, foi possível utilizar as técnicas multivariadas de clusterização para observar outra forma de agrupamento a partir da quantidade de produtos utilizados, além da análise de correspondência entre as variáveis categóricas de porte, sistemas de produção e manejo do solo.

Na análise inicial observou-se que existe uma diferença na proporção de hectares destinados exclusivamente à produção orgânica quando considerado o porte dos produtores. Os produtores com maior propriedade utilizam menos da metade da propriedade para a produção orgânica (com apenas uma exceção). A análise das tabelas de frequência possibilitou observar, também, as diferenças entre os Portes e os sistemas utilizados pelos produtores, de forma que a produção de Grãos é menos utilizada pelos pequenos e minifúndios, mas é utilizada por todos os de Grande porte.

Considerando as formas de manejo de solo, a Manutenção de restos vegetais é mais utilizada dentre os produtores, enquanto que curvas de nível é utilizada por menos da metade. Por fim, Controle Manual é a forma mais comum de manejo de plantas invasoras, utilizada por tordos os produtores de Grande e Minifúndio.

Observando as categorias dos produtos utilizados para nutrição vegetal e manejo de doenças e pragas, notou-se que nenhum tipo de produto é utilizado por todos os produtores. Os produtos considerados como 'Remineralizadores' são os mais utilizados para nutrição vegetal, caracterizado pelos produtos que atuam na incrementação dos minerais no solo.'Controle biológico' é o tipo mais utilizado para o manejo fitossanitário, são os produtos baseados organismos que lidam com fungos e pragas. Além disso, 6 categorias são utilizadas por 50% dos produtores.

Os conglomerados foram gerados a partir do método de Hierárquico de Ward e considerando as distâncias euclidianas. O reagrupamento dos produtores possibilitou o desenvolvimento de grupos baseados na quantidade de produtos utilizados. Foi observado que o conglomerado 3 é caracterizado por produtores que utilizam, em média, 12 tipos de produtos sendo,também, o grupo com menor quantidade de produtores depois do grupo 4 que possui apenas 1. O grupo 2 tem a maior variabilidade na frequência de tipos de produtos utilizados e também é o maior grupo, com 59 produtores.

Por fim, foi possível entender melhor o perfil de cada porte na análise de correspondência. Sem considerar o manejo de plantas invasoras, devido a clareza na análise exploratória, observou-se que os produtores de Grande porte são caracterizados pela prática de produção de Grãos e utilizando Adubação Verde. Os Médios utilizam o

sistema tanto de Fruticultura quanto de Hortaliça fazendo o manejo do solo por rotação de cultura. Os produtores de pequeno porte, são próximos de todas as categorias e ficam ambíguos ou contraditórios à análise apresentada na parte descritiva. Por fim, os produtores do grupo minifúndio são mais próximos de Fruticultura e Manutenção de restos vegetais.

Referências

BRASIL. Lei nº 10.831, de 23 de dezembro de 2003. *Diário Oficial da República Federativa do Brasil*, Brasília, DF, 2003. Disponível em: ⟨https://www.planalto.gov.br/ccivil_03/LEIS/2003/L10.831.htm⟩.

CALINSKI, T.; HARABASZ, J. A dendrite method for cluster analysis. *Communications in Statistics*, Taylor Francis, v. 3, n. 1, p. 1–27, 1974.

CHARRAD, M.; GHAZZALI, N.; BOITEAU, V.; NIKNAFS, A. Nbclust: An r package for determining the relevant number of clusters in a data set. *Journal of Statistical Software*, v. 61, n. 6, p. 1–36, 2014.

CODEPLAN. *AGRICULTURA FAMILIAR NO DISTRITO FEDERAL , DIMENSÕES E DESAFIOS*. [S.l.], 2015.

DUDA, R. O.; HART, P. E. *Pattern classification and scene analysis.* [S.l.]: Wiley, 1973. I-XVII, 1-482 p. (A Wiley-Interscience publication). ISBN 0471223611.

EMATER-DF, G. *A Emater-DF*. 2023. Urlhttps://emater.df.gov.br/a-emater-df/.

FENNER, A. L. D.; ALMEIDA, V. E. S. d.; FRIEDRICH, K.; MILHOMEM, A. P. A. S. Territórios saudáveis e sustentáveis (tss) no distrito federal: agroecologia e impacto dos agrotóxicos. *Saúde em Debate*, Centro Brasileiro de Estudos de Saúde, v. 46, n. spe2, p. 249–261, 2022. ISSN 2358-2898.

GAN, G.; MA, C.; WU, J. *Data Clustering: Theory, Algorithms, and Applications*. Society for Industrial and Applied Mathematics, 2007. (ASA-SIAM Series on Statistics and Applied Probability). ISBN 9780898716238. Disponível em: ⟨https://books.google.com.br/books?id=rlQZAQAAIAAJ⟩.

GORDON, A. *Classification, 2nd Edition*. [S.l.]: CRC Press, 1999. (Chapman & Hall/CRC Monographs on Statistics & Applied Probability). ISBN 9781584888536.

GREENACRE, M. *Correspondence Analysis in Practice*. CRC Press, 2007. (Chapman & Hall/CRC Interdisciplinary Statistics). ISBN 9781420011234. Disponível em: ⟨https://books.google.com.br/books?id=uL6-PKdS0lAC⟩.

HAIR, J.; BLACK, W.; BABIN, B.; ANDERSON, R.; TATHAM, R. *Análise multivariada de dados - 6ed*. Bookman, 2009. ISBN 9788577805341. Disponível em: ⟨https://books.google.com.br/books?id=oFQs_zJI2GwC⟩.

IZENMAN, A. *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*. Springer New York, 2009. (Springer Texts in Statistics). ISBN 9780387781891. Disponível em: ⟨https://books.google.com.br/books?id=1CuznRORa3EC⟩.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. *An Introduction to Statistical Learning: with Applications in R*. Springer New York, 2013. (Springer Texts in Statistics). ISBN 9781461471387. Disponível em: ⟨https://books.google.com.br/books?id=qcI_AAAAQBAJ⟩.

JOHNSON, R.; WICHERN, D. *Applied Multivariate Statistical Analysis*. Pearson Prentice Hall, 2007. (Applied Multivariate Statistical Analysis). ISBN 9780131877153. Disponível em: ⟨https://books.google.com.br/books?id=gFWcQgAACAAJ⟩.

MARCHETTI, F. F.; LOPES, K. C. S. A.; GUYOT, M.; SORRENTINO, M.; LOPES, P. R. Agroecologia: CiÊncia, movimento polÍtico e prÁtica social para mitigaÇÃo e adaptaÇÃo Às mudanÇas climÁticas. *Revista Brasileira de agroecologia*, Associacao Brasileira De Agroecologia, v. 18, n. 1, p. 388–415, 2023. ISSN 1980-9735.

MARDIA, K.; KENT, J.; BIBBY, J. *Multivariate Analysis*. [S.l.]: Academic Press, 1979. (Probability and Mathematical Statistics : a series of monographs and textbooks). ISBN 9780124712508.

MUñOZ, M. S. G.; SOARES, J. P. G.; BRISOLA, M. V.; JUNQUEIRA, A. M. R.; PANTOJA, M. J. Impactos ambientais e socioeconômicos da produção integrada de base ecológica em unidades de produção familiar do distrito federal e entorno. *Revista de Economia e Sociologia Rural*, Sociedade Brasileira de Economia e Sociologia Rural, Edificio Brasilia Radio Center, Brasilia, v. 60, n. 1, p. 1, 2022. ISSN 0103-2003.

PETE, C.; JULIAN, C.; RANDY, K.; THOMAS, K.; THOMAS, R.; COLIN, S.; WIRTH, R. Crisp-dm 1.0. *CRISP-DM Consortium*, 2000.

SABOURIN, E.; SILVA, L. R. T. da; AVILA, M. Lucio de. Construção da política de agroecologia e produção orgânica no distrito federal. *Revista Brasileira de agroecologia*, Associação Brasileira de Agroecologia, v. 14, n. 2, p. 35–50, 2019. ISSN 1980-9735.

SARLE, W. S. *Cubic Clustering Criterion*. SAS Institute, 1983. (SAS technical report). Disponível em: ⟨https://cir.nii.ac.jp/crid/1130000796449908480⟩.

SILVA, C. M. da; BOTELHO, R. V.; FARIA, C. M. D. R. Ação de extratos de cinamomo sobre. *Bioscience journal*, Universidade Federal de Uberlândia, v. 30, 2014. ISSN 1981-3163.