PCA:

Uma Ferramenta Matemática para a Análise dos COREDEs Agropecuários do Rio Grande do Sul

Rafael Pentiado Poerschke

TCC B - Orientação: Prof. Dr. João Roberto Lazzarin [1]

[1]Centro de Ciências Naturais e Exatas (CCNE)
Universidade Federal de Santa Maria (UFSM)

26 de julho de 2024

Roteiro

Referências

Referências Principais

- JOLLIFFE, Ian T. (1990). Principal component analysis: a beginner's guide—I. Introduction and application. Weather, v. 45, n. 10, p. 375-382.
- JOHNSON, R. A.; WICHERN, D. W. and others. (2002) *Applied multivariate statistical analysis*, Prentice hall Upper Saddle River, NJ.
- JÖRESKOG, K. G. Basic ideas of factor and component analysis. Advances in factor analysis and structural equation models, [S.l.], p.5–20, 1979.

Referências Complementares

- POOLE, D. Linear Algebra: a modern introduction. [S.l.]: CENGAGE Learning, Boston MA, 2011.
- MARDIA KANTI V.; KENT, J.; BIBBY, J. M. Multivariate Analysis. [S.l.]: Academic Press, 1st edition, 1979.
- VIDAL, R.; MA, Y.; SASTRY, S. Generalized principal component analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, [S.l.], v.27, n.12, p.1–15, 2005.

Roteiro

Componentes Principais: Ideia principal/motivação

Dados em elevada dimensão:

- Histórico das 400 empresas na B3
  - Categorias?
- Imagens: 9x13 cm
  - Matriz menor (729x553)?
- Histórico de buscas na Internet
  - *Ranqueamento* de páginas.

Componentes Principais: Motivação

Dados em elevada dimensão:

- Histórico das 400 Empresas na B3
  - Categorias?
- Imagens: 9x13 cm
  - Matriz menor? Sistema RGB: (729x553)x3.
- Histórico de buscas na Internet
  - *Ranqueamento* de páginas.

Número de Componentes Principais

Figura: 3 Componentes

Número de Componentes Principais

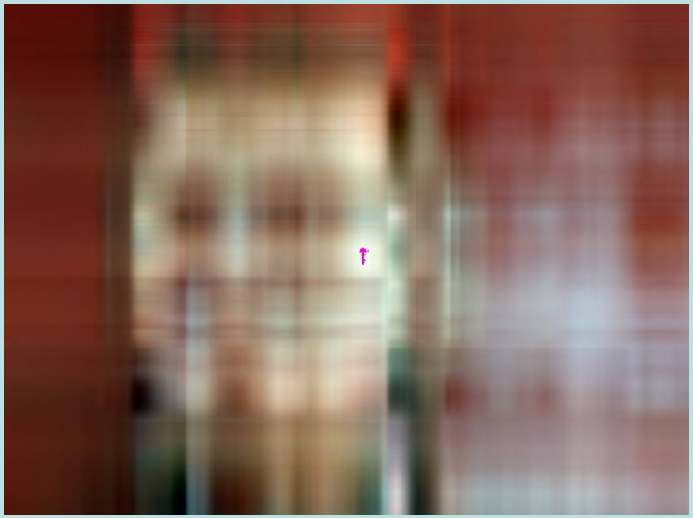


Figura: 3 Componentes

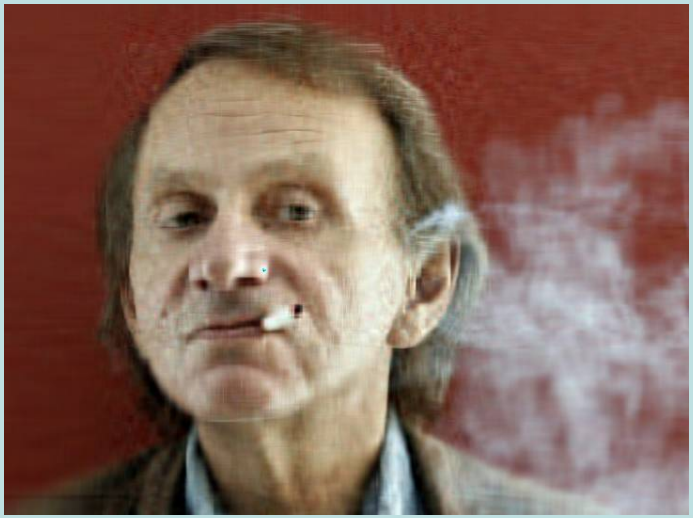


Figura: 29 Componentes

Número de Componentes Principais

Figura: 100 Componentes

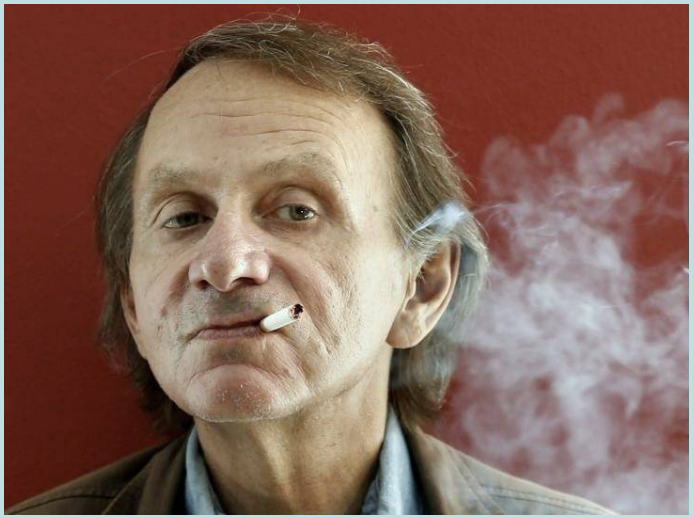


Figura: 291 Componentes

Número de Componentes Principais

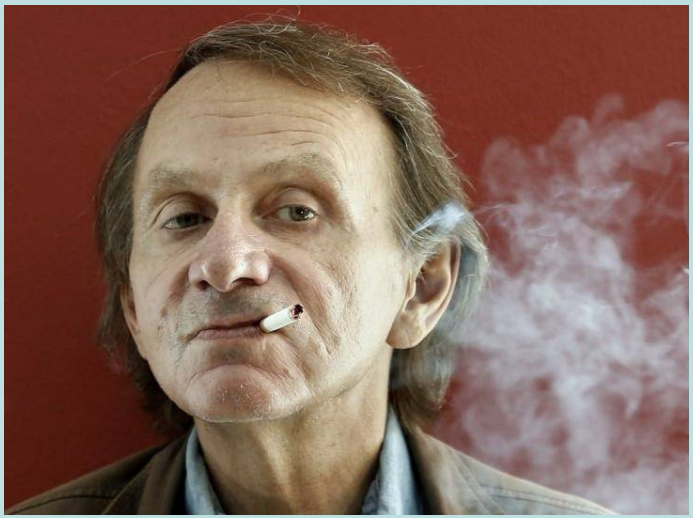


Figura: 291 Componentes

Figura: Original (729cols)

Questões centrais

- **Análise de Componentes Principais**: O que é? De onde vem? Como funciona?
- **Decomposição de Matrizes**: Qual tipo?
- **Economia do Desenvolvimento**: Aplicação da PCA.

Questões centrais

- **Análise de Componentes Principais**: O que é? De onde vem? Como funciona?
- **Decomposição de Matrizes**: Qual tipo?
- **Economia do Desenvolvimento**: Aplicação da PCA.

Questões centrais

- **Análise de Componentes Principais**: O que é? De onde vem? Como funciona?
- **Decomposição de Matrizes**: Qual tipo?
- **Economia do Desenvolvimento**: Aplicação da PCA.

Objetivos de Pesquisa

- **Apresentar** com rigor a matemática por trás da técnica da PCA.
- **Ilustrar** o método com uma aplicação da PCA a fim de selecionar os componentes principais de uma base de dados para o Rio Grande do Sul (RS).
- **Investigar** os COREDEs agropecuários gaúchos, considerando a existência e o grau de similaridade entre os municípios com base nos dados do Censo Agropecuário de 2017.

Aplicação: Problema de Pesquisa

O foco do presente estudo está em aplicar uma transformação linear que leve a redução da dimensão inicial de dados. Especificamente, intenta-se selecionar os **Componentes Principais**, com a utilização da linguagem estatística R a fim de investigar os COREDEs agropecuários gaúchos, considerando a existência de similaridade entre os municípios com base nos dados do Censo Agropecuário de 2017.

Aplicação: Problema de Pesquisa

Em um universo de 127 municípios e 15 variáveis, agregados em 8 COREDEs predominantemente **agropecuários**, questiona-se o quão homogêneo será esse grupo, isto é, **em que medida** a agregação por contiguidade, garantiria a homogeneidade dos COREDEs.

Hipótese

A agregação de um grupo de municípios no estado do RS por **contiguidade** - na forma dos COREDEs - não é suficiente para garantir a homogeneidade entre os municípios que fazem parte dos COREDEs agropecuários.

Roteiro

Análise de Componentes Principais

- **O que é** (para a Matemática): é uma transformação linear;
  - **Ingredientes**: ortogonalidade e decomposição em autovalores.
- **De onde vem**: Pearson (1901) e Hotelling (1933);
- **Para que serve**: possibilita a redução da dimensão inicial de dados (colunas).
- **Como faz**: APC é um modelo linear - uma combinação linear.

Análise de Componentes Principais

- **O que é** (para a Matemática): é uma transformação linear;
  - **Ingredientes**: ortogonalidade e decomposição em autovalores.
- **De onde vem**: Pearson (1901) e Hotelling (1933);
- **Para que serve**: possibilita a redução da dimensão inicial de dados (colunas).
- **Como faz**: APC é um modelo linear - uma combinação linear.

Análise de Componentes Principais

- **O que é** (para a Matemática): é uma transformação linear;
  - **Ingredientes**: ortogonalidade e decomposição em autovalores.
- **De onde vem**: Pearson (1901) e Hotelling (1933);
- **Para que serve**: possibilita a redução da dimensão inicial de dados (colunas).
- **Como faz**: APC é um modelo linear - uma combinação linear.

Análise de Componentes Principais

- **O que é** (para a Matemática): é uma transformação linear;
  - **Ingredientes**: ortogonalidade e decomposição em autovalores.
- **De onde vem**: Pearson (1901) e Hotelling (1933);
- **Para que serve**: possibilita a redução da dimensão inicial de dados (colunas).
- **Como faz**: APC é um modelo linear - uma combinação linear.

Análise de Componentes Principais

- **O que é** (para a Matemática): é uma transformação linear;
  - **Ingredientes**: ortogonalidade e decomposição em autovalores.
- **De onde vem**: Pearson (1901) e Hotelling (1933);
- **Para que serve**: possibilita a redução da dimensão inicial de dados (colunas).
- **Como faz**: APC é um modelo linear - uma combinação linear.

Componentes Principais

A Análise de Componentes Principais é um problema no qual busca-se estimar um subespaço de dimensão inferior m de um conjunto de pontos em um espaço de dimensão maior $\mathbb{R}^p$ dispostos em uma matriz $\mathbf{X}_{(n \times p)} = [\mathbf{x}_1\ \mathbf{x}_2\ \ldots\ \mathbf{x}_p]$ formada por p variáveis aleatórias correlacionadas entre si.

Componentes Principais

Esse problema pode ser modelado como uma questão **estatística** ou **geométrica**. Existe uma terceira abordagem, no qual ACP é vista como um problema de **aproximação** de uma matriz de **menor posto** em relação original.

Gráfico de Dispersão: duas variáveis aleatórias

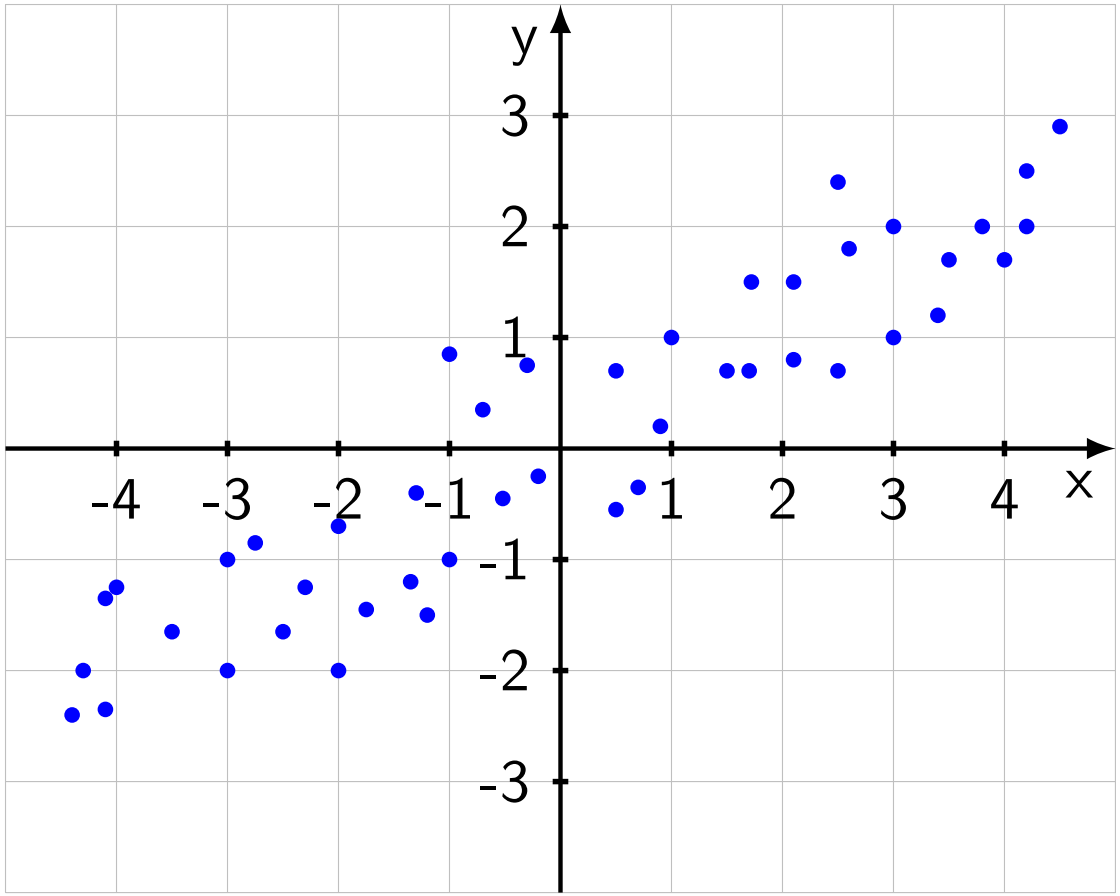
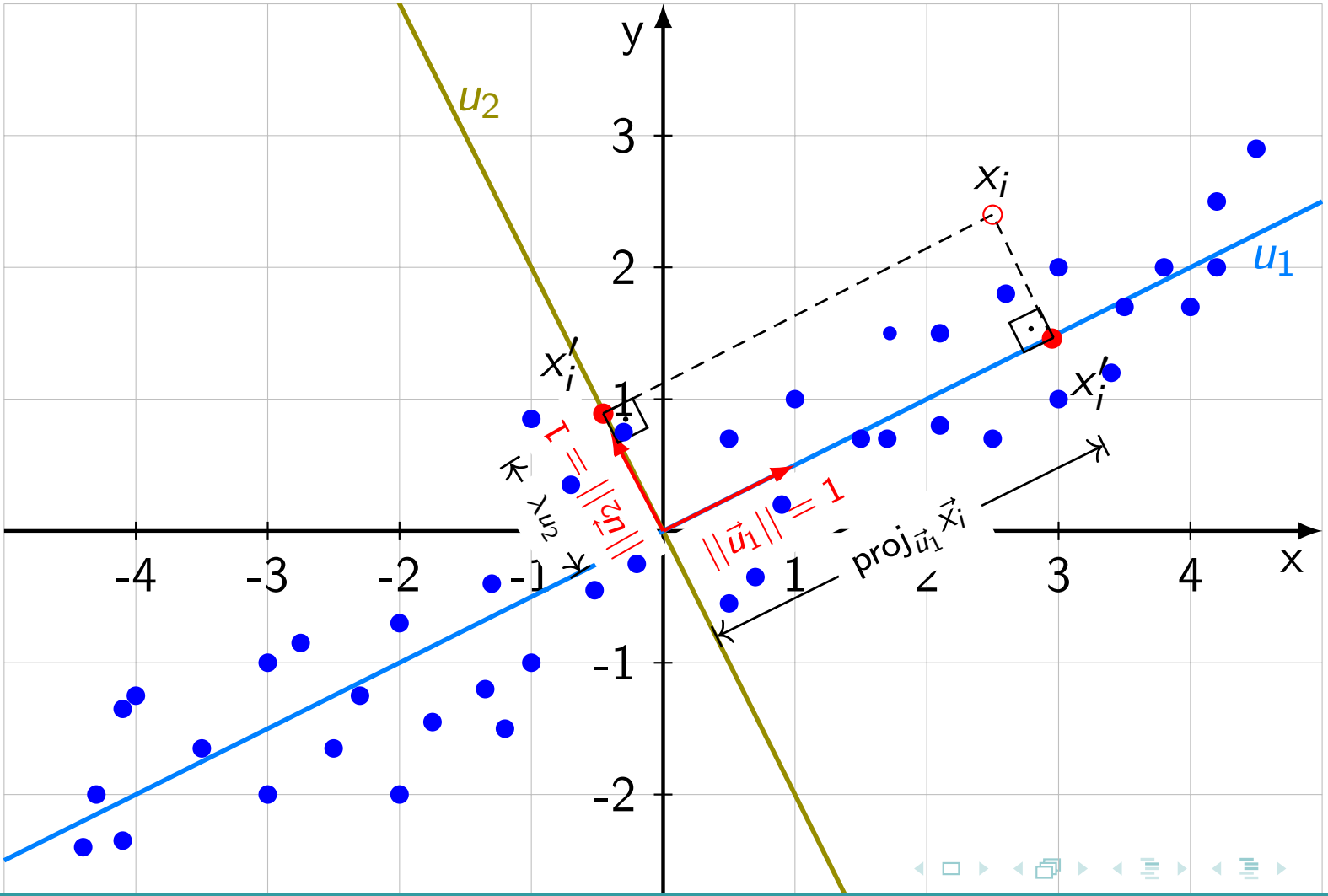


Figura: Eixos Coordenados

O Problema da ACP



Componentes Principais

Teorema: Componentes Principais de Variáveis Aleatórias[1]

Assuma que posto $(\mathbf{S}_X) = p$. Então os p componentes principais de uma variável aleatória multivariada $\mathbf{X} \in \mathbb{R}^p$, denotados por $\mathbf{w}_j$ para $j = 1, 2, \ldots, p$, são dados por

$$\mathbf{w}_j = \mathbf{u}_j^T \mathbf{X},$$

onde $\mathbf{u} \in \mathbb{R}^p$ e $\{\mathbf{u}_j\}_{j=1}^p$ são os p autovetores de $\mathbf{S}_X$ associados aos maiores autovalores λ_j. Além disso, $\lambda_j = \text{Var}(\mathbf{w}_j)$ para $j = 1, 2, \ldots, p$.

[1] A demonstração do teorema pode ser consultada em Jolliffe (1990).

Componentes Principais

Definimos cada novo $\mathbf{w}_j$, com dimensão $(n \times 1)$, em função linear dos autovetores de $\mathbf{S}$, combinados com os vetores que compõe $\mathbf{X}$ do seguinte modo

$$\mathbf{w}_j = \mathbf{u}_j^T \mathbf{X} = u_{j1}\mathbf{x}_1 + u_{j2}\mathbf{x}_2 + \cdots + u_{jp}\mathbf{x}_p, \qquad \forall j = 1, 2, \ldots, p. \quad (1)$$

Por exemplo, $\mathbf{u}_1$ é um vetor dado por $\mathbf{u}_1 = [u_{11}\ u_{12}\ \ldots\ u_{1p}]$. Portanto, o primeiro componente principal será a combinação linear

$$\mathbf{w}_1 = \mathbf{u}_1^T \mathbf{X} = u_{11}\mathbf{x}_1 + u_{12}\mathbf{x}_2 + \cdots + u_{1p}\mathbf{x}_p\,.$$

Componentes Principais

O segundo componente principal será a combinação linear

$$\mathbf{w}_2 = \mathbf{u}_2^T \mathbf{X} = u_{21}\mathbf{x}_1 + u_{22}\mathbf{x}_2 + \cdots + u_{2p}\mathbf{x}_p \ .$$

Assim, o primeiro componente principal $\mathbf{w}_1^T = [w_{11}\ w_{12}\ \ldots\ w_{1n}]$ tem coordenadas dadas pela combinação

$$\begin{aligned}
w_{11} &= u_{11}x_{11} + u_{12}x_{12} + \cdots + u_{1p}x_{1p};\\
w_{12} &= u_{11}x_{21} + u_{12}x_{22} + \cdots + u_{1p}x_{2p};\\
\vdots \ &= \qquad\qquad\qquad\qquad \vdots\\
w_{1n} &= u_{11}x_{n1} + u_{12}x_{n2} + \cdots + u_{1p}x_{np}\,.
\end{aligned}$$

Problema:

$$\text{Var}(\mathbf{X}^T\mathbf{u}_k) = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i^T\mathbf{u}_k)^2 = \mathbf{u}^T\frac{\mathbf{X}^T\mathbf{X}}{n}\mathbf{u} = \mathbf{u}^T\mathbf{S}\mathbf{u} = \text{Var}(\mathbf{w}_i), \tag{2}$$

Qual a direção que maximiza a variância dos dados?

Solução: para a primeira direção

Temos o seguinte problema de maximização de (2):

$$\begin{cases} \max\limits_{\mathbf{u}} & (\mathbf{u}^T \mathbf{S}\mathbf{u}) \\ \text{sujeito a:} & ||\mathbf{u}|| = \mathbf{u}^T\mathbf{u} = 1. \end{cases}$$

Resposta: Multiplicadores de Lagrange.

$$\nabla f(\mathbf{x}_0) = \alpha_0 \nabla g_0(\mathbf{x}_0).$$

Solução: para a primeira direção

Para o primeiro componente (direção), temos que maximizar

$$\begin{cases} f(\mathbf{u}) = & \mathbf{u}^T\mathbf{S}\mathbf{u} \\ \text{sujeito a:} & g_0(\mathbf{u}) = \mathbf{u}^T\mathbf{u} - \mathbf{1} = 0 \end{cases}$$

Aplicando os limites, temos

$$\nabla f(\mathbf{u}) = 2\mathbf{S}\mathbf{u};$$

$$\nabla g_0(\mathbf{u}) = 2\mathbf{u}.$$

tais que $\nabla f(\mathbf{u}) = \alpha_0 \nabla g_0(\mathbf{u})$, ou seja,

$$2\mathbf{S}\mathbf{u} = 2\alpha_0\mathbf{u} \Leftrightarrow (\mathbf{S} - \alpha_0 I)\mathbf{u} = 0.$$

Solução: para a primeira direção

Se

$$2\mathbf{S}\mathbf{u} = 2\alpha_0\mathbf{u} \Leftrightarrow (\mathbf{S} - \alpha_0 I)\mathbf{u} = 0,$$

logo as soluções procuradas necessariamente são **autovetores**. Tomando $\mathbf{u} = \mathbf{u}_1$, como sendo um autovalor associado **ao maior autovalor** dentre todos, garantimos uma solução com a maior variância possível para nosso problema.

Solução: para a p-ésima direção

Temos o seguinte problema de maximização de

$$\begin{cases} f(\mathbf{u}) = & \mathbf{u}^T\mathbf{S}\mathbf{u} \\ \\ \text{com restrições:} & \begin{cases} g_0(\mathbf{u}) = \mathbf{u}^T\mathbf{u} - \mathbf{1} = 0 \\ g_1(\mathbf{u}) = \mathbf{u}^T\mathbf{u_1} = 0 \\ \vdots \\ g_{i-1}(\mathbf{u}) = \mathbf{u}^T\mathbf{u_{i-1}} = 0 \end{cases} \end{cases}$$

$$\nabla f(\mathbf{x}_0) = \alpha_0 \nabla g_0(\mathbf{x}_0) + \alpha_1 \nabla g_1(\mathbf{x}_0) + \cdots + \alpha_h \nabla g_h(\mathbf{x}_0).$$

Componentes Principais: truncamento

Corolário: Redefinindo os Componentes Principais de Variáveis Aleatórias

Seja $m \leq p$. Assuma que posto $(\mathbf{S}_X) \geq m$. Então os primeiros m componentes principais de uma variável aleatória multivariada $\mathbf{X} \in \mathbb{R}^p$ são dados por $\mathbf{w}_j = \mathbf{u}_j^T \mathbf{X}$, onde $\mathbf{u} \in \mathbb{R}^p$ e $\{\mathbf{u}_i\}_{i=1}^m$ são os m autovetores de $\mathbf{S}_X$ associados aos maiores autovalores $\lambda_i > 0$.

Roteiro

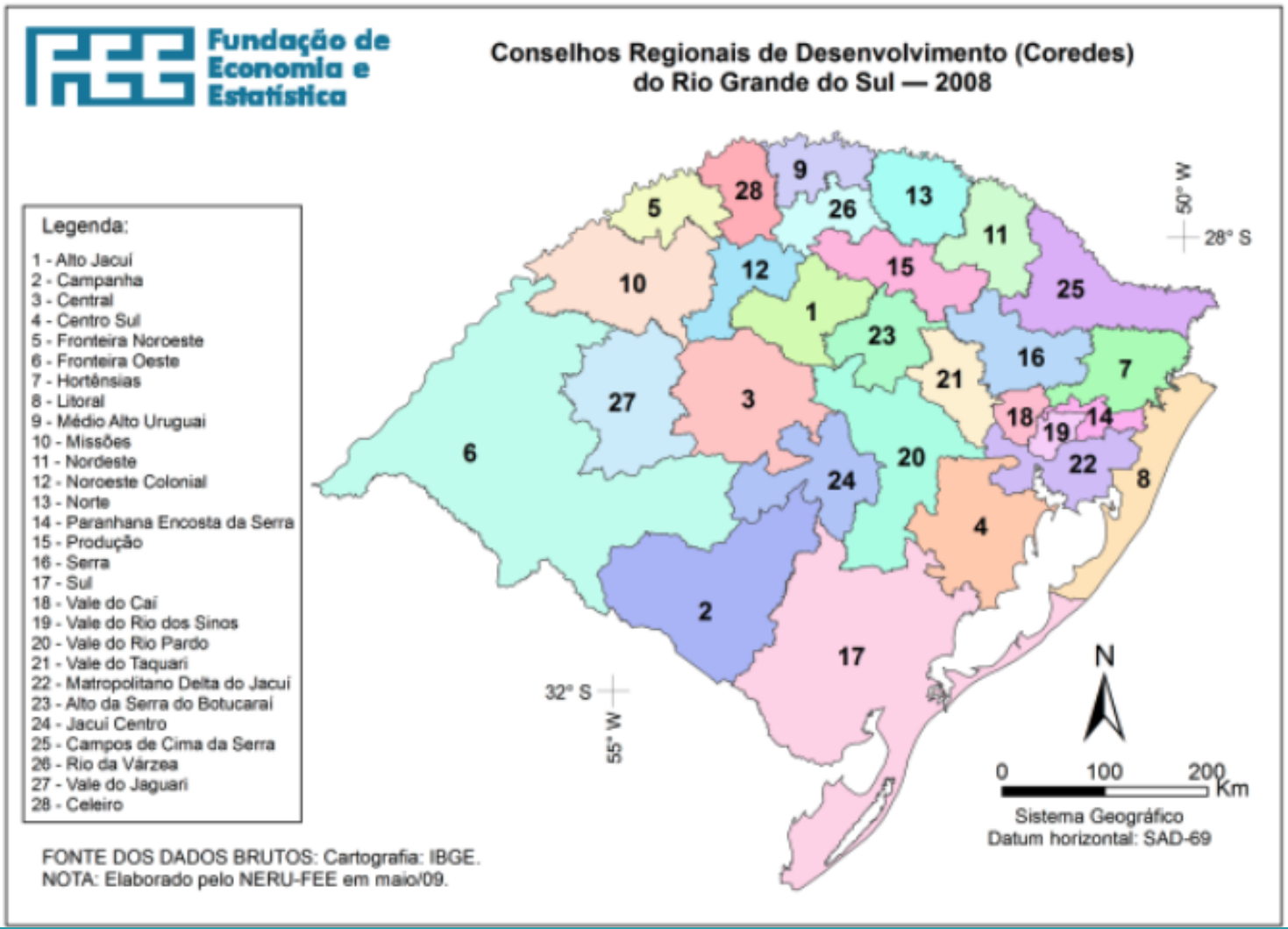
COREDEs: Origem

Definição: Conselhos Regionais de Desenvolvimento

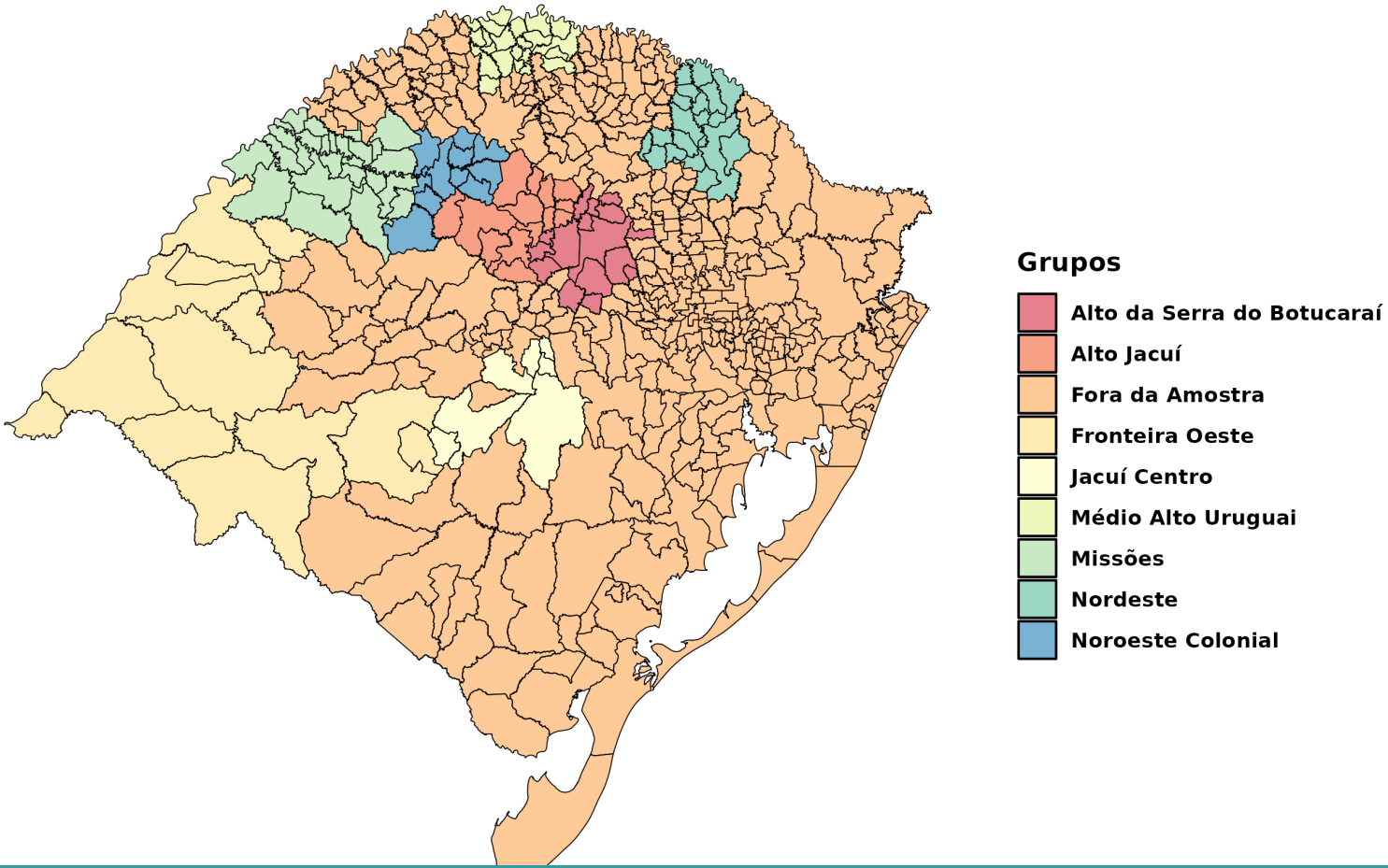
Os Conselhos Regionais de Desenvolvimento – COREDEs, criados oficialmente pela Lei 10.283 de 17 de outubro de 1994, são um fórum de discussão para a promoção de políticas e ações que visam o desenvolvimento regional.

Temos **28** COREDEs e **497** municípios no estado - começou com 21.

COREDEs: Localização



COREDEs Agropecuários



Dados utilizados

Num conjunto de 127 municípios, agregados em 8 COREDEs predominantemente agropecuários com 15 variáveis, sendo todas coletadas no IBGE e listadas no Quadro 1.

Os dados utilizados na pesquisa refletem a proporção do município em relação ao total da amostra, seja a variável medida em unidades, pessoas, unidades monetárias etc.

As variáveis Área Relativa, que mede a relação entre área explorada e a área total do município, bem como Índice de Desenvolvimento Socioeconômico (IDESE) dos municípios do Rio Grande do Sul foram descartadas.

Dados utilizados

Sigla	Nome da Variável	Referência	Unidade de Medida	Fonte
fin_veg	Financiamento (Prod. Vegetal)	Tabela 6895	N. de Estabelecimentos	IBGE
fin_pec	Financiamento (Prod. Pecuária)	Tabela 6895	N. de Estabelecimentos	IBGE
ass_veg	Assistência Técnica (Prod. Vegetal)	Tabela 6844	N. de Estabelecimentos	IBGE
ass_pec	Assistência Técnica (Prod. Pecuária)	Tabela 6844	N. de Estabelecimentos	IBGE
colhe	Colheitadeiras	Tabela 6874	Unidades	IBGE
trat	Tratores	Tabela 6869	Unidades	IBGE
gado	Rebanho Bovino	Tabela 6907	Rebanho	IBGE
pea	População Economicamente Ativa	Tabela 6887	Pessoas	IBGE
pop	População Residente	Tabela 6579	Pessoas	IBGE
rec_veg	Receitas com Lavouras	Tabela 6897	Mil R$	IBGE
val_pec	Valor da Produção Pecuária	Tabela 6898	Mil R$	IBGE
val_veg	Valor da Produção Vegetal	Tabela 6897	Mil R$	IBGE
irriga	Irrigação	Tabela 6857	N. de Estabelecimentos	IBGE
adubo	Adubação	Tabela 6847	N. de Estabelecimentos	IBGE
area_rela	Área Explorada/Área Total	15761**	Área (km^2)	IBGE
area_exp	Área Total Explorada	Tabela 6878	Área (ha)	IBGE
idese	IDESE	Bloco Renda	Numero Índice	FEE***

* - Os dados são referentes ao Censo Agropecuário 2017, exceto pela Área Total dos Municípios e IDESE Bloco Renda.
** - Áreas Territoriais (Instituto Brasileiro de Geografia e Estatística).
*** - Fundação de Economia e Estatística (FEE).

Tabela: Variáveis utilizadas*

Resultados

Correlação das Variáveis com os Autovetores			
Autovalores	$\lambda_1^R = 8,98$	$\lambda_2^R = 2,62$	$\lambda_3^R = 1,47$
	Autovetor 1	**Autovetor 2**	**Autovetor 3**
val_veg	0,277	0,000	0,421
fin_veg	0,143	-0,507	-0,008
rec_veg	0,277	0,019	0,412
ass_veg	0,149	-0,509	-0,201
fin_pec	0,239	0,147	-0,418
val_pec	0,289	0,245	-0,135
gado	0,277	0,303	-0,119
ass_pec	0,284	0,203	-0,280
adubo	0,214	-0,390	-0,328
colhe	0,267	-0,169	0,353
trat	0,308	-0,124	0,169
pea	0,311	-0,087	-0,209
pop	0,280	0,020	0,141
area_exp	0,297	0,249	0,028
irriga	0,174	0,053	-0,042
area_rela	—	—	—
idese	—	—	—

Tabela: Correlação das Variáveis com os Autovetores (Matriz $\mathbf{R}_X$)

Resultados: da variância

Juntos, os **três primeiros autovalores** responderam por cerca de 87% da variância do conjunto original de dados.

A proporção explicada da variância original é a soma dos autovalores dos componentes retidos dividido pelo traço da matriz no qual os autovalores foram extraídos:
Total Explicado $= \frac{13,07}{15} = 0,8715$.

Resultados: relação estabelecida entre as variáveis originais com os autovetores

É possível verificar que **todo o conjunto das variáveis** tem uma relação diretamente proporcional com o **Autovetor 1**.

As variáveis Número de Tratores (0,308) e População Economicamente Ativa (0,311) apresentam as maiores magnitudes.

Resultados: relação estabelecida entre as variáveis originais com os autovetores

As variáveis financeiras relacionadas à agricultura denotam maior correlação com o **Autovetor 3**.

As maiores correlações entre variáveis e autovetor verificam-se no Valor da Produção Vegetal (0,421), as Receitas da Produção Vegetal (0,412), seguidos do Número de Colheitadeiras (0,353) e Número de Tratores (0,169).

Esse comportamento indica que o **Autovetor 3** representa, em sua maioria, a variabilidade das **variáveis financeiras** que mantêm relação com a **produção agrícol**a e, em especial, intensivas no uso de máquinas e implementos agrícolas.

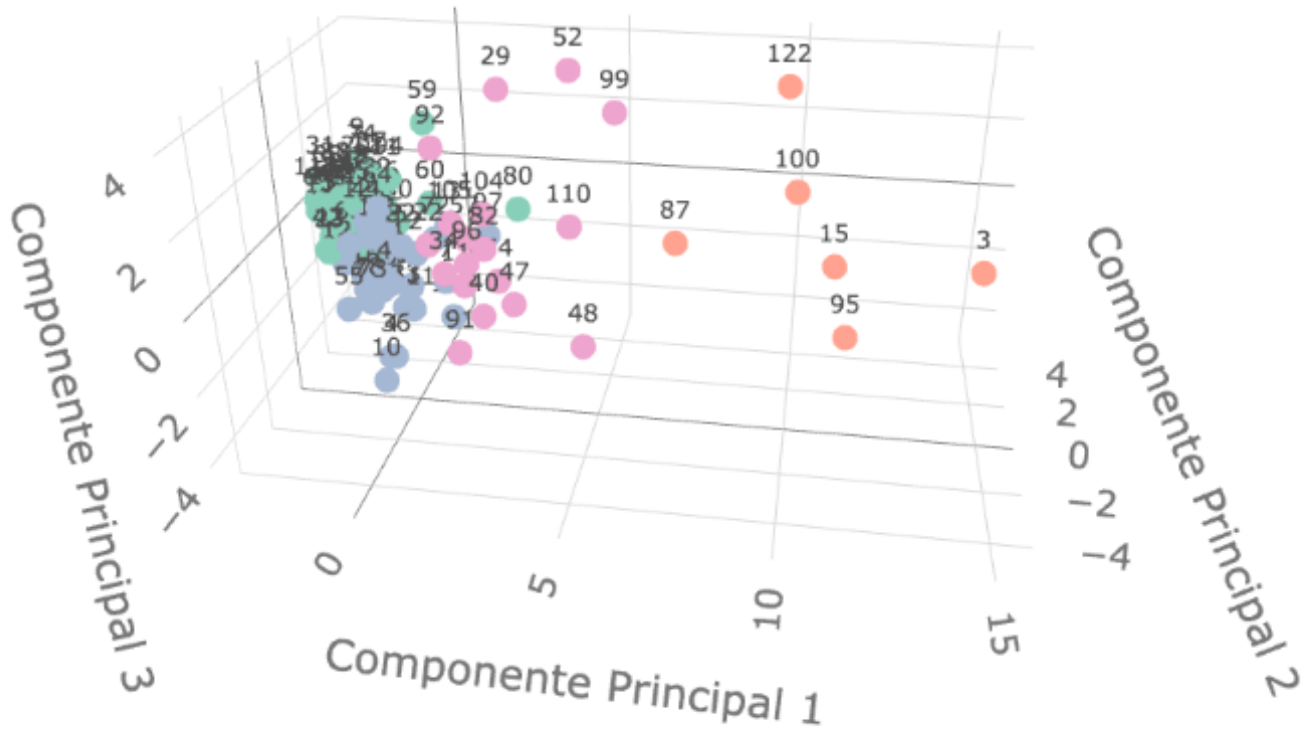
Resultados: relação estabelecida entre as variáveis originais com os autovetores

As variáveis relacionadas à **produção pecuária** tiveram maior relação com o **Autovetor 2**.

As variáveis com relação direta foram o Rebanho Bovino (0,303), o Valor da Produção Pecuária (0,245) e o Financiamento da Pecuária (0,147).

Já as variáveis financeiras ligadas à **produção agrícola** obtiveram magnitude significativa mas com **sinais opostos**. Isso indica que o **Autovetor 2** tem uma relação direta com as variáveis ligadas à produção **pecuária**.

Grupos - COREDEs Agropecuários em três dimensões



Grupo 2

Código	Município	Componente 1	Componente 2	Componente 3
3	Alegrete	15,39	5,23	-2,64
15	Cachoeira do Sul	10,94	-3,24	1,47
87	Rosário do Sul	6,60	3,46	-1,47
95	Santana do Livramento	11,57	4,71	-4,82
100	São Gabriel	9,98	0,79	1,46
121	Uruguaiana	9,59	2,78	3,53

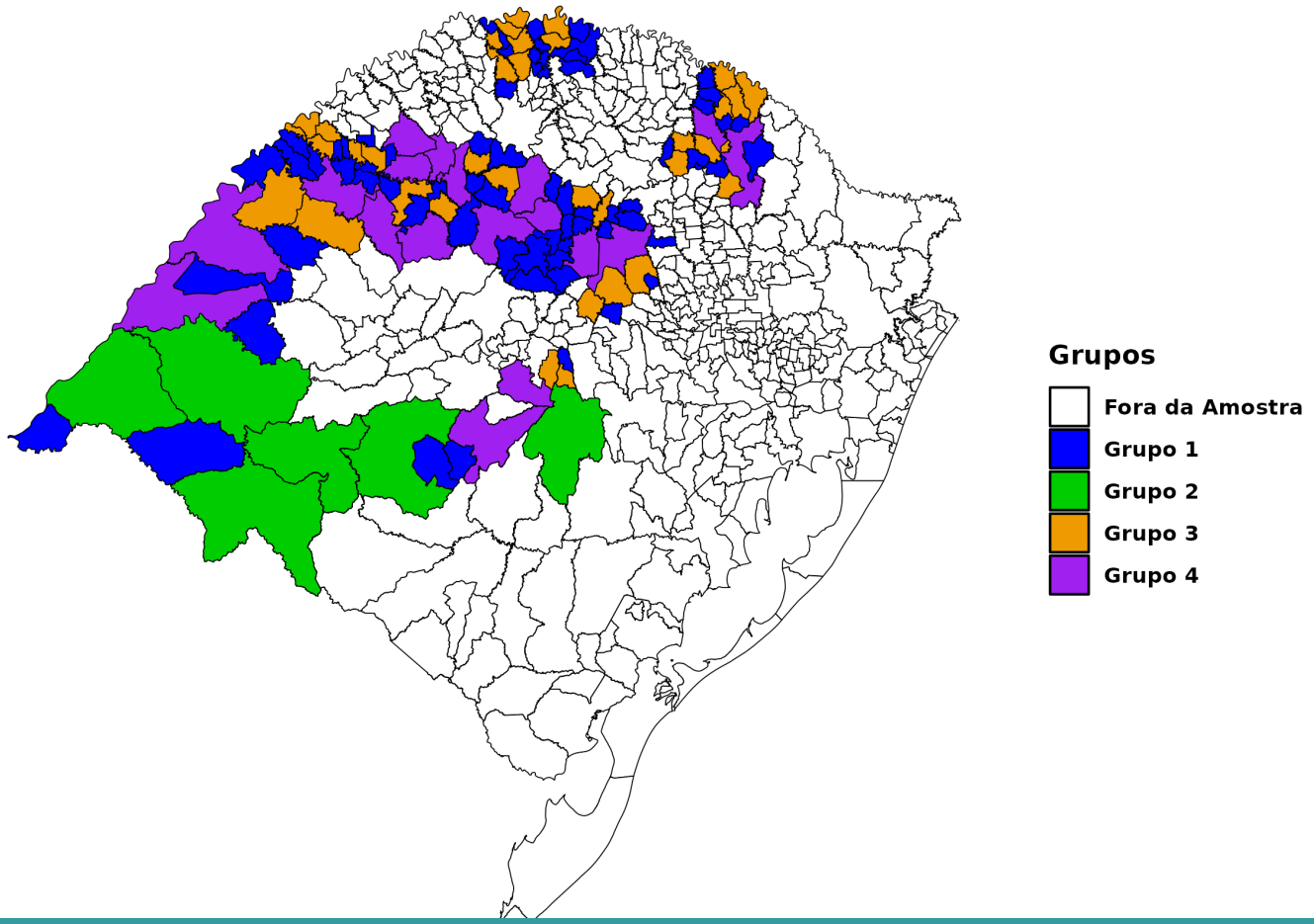
Tabela: Escores dos municípios do Grupo 2

Média dos Escores por Grupos (de uma matriz 127 × 15)

Municípios Agrupados	Grupo	Autovetor 1	Autovetor 2	Autovetor 3
69	1	-1,62	0,71	0,15
6	2	10,68	2,29	-0,41
34	3	-011	-0,91	-0,75
18	4	2,81	2,86	-1,75

Tabela: Média dos Escores dos Municípios de cada Grupo

Grupos - COREDEs Agropecuários



Roteiro

Com os **três autovetores** estimados, foi possível a identificação de **quatro agrupamentos** potenciais de municípios dentro dos COREDEs, e esse resultado tem implicações práticas significativas. Essa segmentação pode servir como uma ferramenta estratégica para políticas agrícolas e de desenvolvimento regional, permitindo a adaptação de estratégias específicas às características distintas de cada grupo.

Mostramos que ainda assim, **dentro** de alguns COREDEs existe certa **heterogeneidade**

PCA:

Uma Ferramenta Matemática para a Análise dos COREDEs Agropecuários do Rio Grande do Sul

Rafael Pentiado Poerschke

TCC B - Orientação: Prof. Dr. João Roberto Lazzarin [1]

[1]Centro de Ciências Naturais e Exatas (CCNE)
Universidade Federal de Santa Maria (UFSM)

26 de julho de 2024