Universidade Federal do Paraná
Setor de Ciências Exatas
Departamento de Estatística
Programa de Especialização em *Data Science* e *Big Data*

Adriana Moraes de Carvalho

Algoritmos de agrupamento para determinar grupos de Unidades Judiciárias semelhantes conforme Resolução 219/2016 do Conselho Nacional de Justiça

São Paulo
2022

Adriana Moraes de Carvalho

Algoritmos de agrupamento para determinar grupos de Unidades Judiciárias semelhantes conforme Resolução 219/2016 do Conselho Nacional de Justiça

Monografia apresentada ao Programa de Especialização em Data Science e Big Data da Universidade Federal do Paraná como requisito parcial para a obtenção do grau de especialista.

Orientador: Prof. Dr. Fernando Mayer

São Paulo
2022

Especialização em Data Science e Big Data
Universidade Federal do Paraná
dsbd.leg.ufpr.br

Algoritmos de agrupamento para determinar grupos de Unidades Judiciárias semelhantes conforme Resolução 219/2016 do Conselho Nacional de Justiça

Adriana Moraes de Carvalho[1]
Fernando Mayer[2]

Resumo

O Conselho Nacional de Justiça (CNJ) por meio da Resolução nº 219 de 26/04/2016, estabelece critérios para o quantitativo mínimo de servidores nas unidades judiciárias de 1º e 2º graus, chamada de lotação paradigma. Para definição da lotação paradigma, é necessário inicialmente agrupar as unidades judiciárias por critério de semelhança, relacionados à competência material, base territorial, entrância ou outro critério objetivo a ser observado. Pensando em estabelecer um método de agrupamento que considere analisar conjuntamente todos os critérios estabelecidos na referida Resolução, é que este estudo se propõe em testar algoritmos de agrupamento para determinar quais grupos de unidades judiciárias são semelhantes entre si. Os resultados encontrados neste estudo se apresentaram bastante coerentes, à medida que foram obtidos grupos de unidades judiciárias diferenciadas por tipo de Entrância e Competência Predominante, e, além disso, foi possível diferenciar unidades com características bem peculiares em seus dados, a exemplo da Vara de Execução Penal e de alguns Juizados Especiais do Tribunal de Justiça do Amapá.

Palavras-chave: cluster, análise de componentes principais, unidades judiciárias.

Abstract

The National Council of Justice (CNJ) through Resolution 219 of April 26rd 2016, establishes criteria for the minimum number of servers in 1st and 2nd-degree judicial units, called paradigm capacity. For the definition of the paradigm capacity, it is initially necessary to group the judicial units by similarity criteria, related to material competence, territorial basis, court or other objective criteria to be observed. Thinking of establishing a clustering method that considers analyzing all the criteria established in the Resolution together, this study proposes to test clustering algorithms to determine which groups of judicial units are similar to each other. The results found in this study were quite coherent, as groups of judicial units were obtained, differentiated by the type of Entrance and Predominant Jurisdiction, and, besides this, it was possible to differentiate units with very peculiar characteristics in their data, such as the Court of Criminal Execution and some Special Courts of the Amapá Court of Justice.

Keywords: cluster, principal components analysis, judicial units.

1 Introdução

A Resolução nº 219/2016 é um dos elementos-chave para promover a Política Nacional de Atenção Prioritária ao Primeiro Grau de Jurisdição, que tem como objetivo desenvolver e implementar medidas concretas e permanentes para melhorar os serviços judiciais prestados pela primeira instância dos tribunais brasileiros.

A referida Resolução entra neste contexto com uma proposta de estabelecer parâmetros objetivos para a distribuição de força trabalho entre os graus de jurisdição e entre unidades judiciais do mesmo nível. Nela, há critérios que estabelecem o quantitativo mínimo de servidores das unidades judiciárias semelhantes.

Para o agrupamento, de acordo com a Resolução, podem ser adotados critérios de semelhança como competência material, base territorial, entrância ou outro parâmetro objetivo.

A forma como o agrupamento é feito impacta diretamente nos resultados da lotação paradigma, já que para o cálculo são comparados dados de unidades judiciárias pertencentes ao mesmo grupo.

Dada a importância deste tema dentro do judiciário e as questões complexas que envolvem diversos pontos da Resolução nº 219/2016, especialmente na dificuldade em se agrupar as unidades judiciais por critério de semelhança, é que este estudo apresenta propostas de

[1]Aluno do programa de Especialização em Data Science & Big Data, adriana.moraesc@gmail.com.
[2] Professor do Departamento de Estatística - DEST/UFPR.

usar algoritmos para realizar o agrupamento.

Para alcançar o objetivo deste estudo, serão usadas técnicas de análise de agrupamentos cuja finalidade é alocar as observações em grupos, e definir a quantidade de grupos que deseja formar, a partir da verificação da existência de comportamentos semelhantes entre observações em relação a determinadas variáveis (Favero & Belfiore, 2017).

2 Materiais e Métodos

A primeira etapa do trabalho consistiu em elencar variáveis e indicadores relacionados a gestão processual, produtividade, tempo do processo, densidade demográfica, tipo de entrância e a competência predominante das 59 unidades judiciárias do 1º grau, do Tribunal de Justiça do Estado do Amapá (TJAP).

As variáveis disponíveis neste estudo são de tipos mistos, e por este motivo é importante a definição do tipo de medida para cada uma delas. A Tabela 1 descreve as variáveis e indicadores utilizadas neste estudo:

Tabela1: Descrição das variáveis/indicadores

Variável/ indicador	Descrição
Processos Novos (Cn)	Total de casos novos que ingressaram ou foram protocolizados durante o período base
Processos Julgados (Sent)	Total de processos que foram julgados durante o período base
IPSJud	Média de processos baixados por servidor durante o período base
Processos Pendentes líquidos (Tpl)	Todos os processos que não tiveram movimentos de baixa até final do período base, desconsiderados os processos suspensos, sobrestados ou em arquivo provisório
Decisões (Dec)	Total de decisões interlocutórias proferidas durante o período base
Audiências (Aud)	Total de audiências realizadas no processo durante o período o período base
Audiências Conciliatórias (AudConc)	Total de audiências de conciliação, de mediação e do art. 334 do CPC realizadas
Taxa de Congestionamento Líquido (Tcl)	Percentual de casos que permaneceram pendentes de solução ao final do período base, desconsiderados os processos suspensos, sobrestados ou em arquivo provisório
Índice de atendimento à demanda (Iad)	Média de processos baixados por caso novo durante o período o período base
TpSent	Tempo médio (dias) entre o início do processo e o primeiro julgamento
TpBaix	Tempo médio (dias) entre o início do processo e a primeira baixa
TpCpL	Tempo médio (dias) decorrido entre a o início da ação judicial e o último dia do ano base
hab/km2	Variação da densidade demográfica
Classificação da unidade judiciária (Classe)	Classifica a unidade judiciária de acordo com a competência predominante (no total teremos 18 tipos de classificação)
Entrância	Indica se é entrância inicial ou final

Fonte: Resolução 76/2009 - CNJ

As variáveis Classe e Entrância são do tipo categóricas nominas, e novas variáveis serão criadas para quantificar essas características.

A variável Classe representa a classificação das unidades judiciárias de acordo com a descrição de suas competências predominantes, sendo elas: Cível, Fazenda Pública, Família, Órfãos e Sucessões, Criminal, Tribunal do Júri, Juizado Especial Criminal, Execução Penal, Execução Penal Medidas Alternativas, Infância E Juventude, Violência Doméstica e Familiar Contra a Mulher, Juizado Especial Cível, Juizado Especial da Fazenda Pública, Juízo Único, Juizado Especial Cível e Criminal, Recursos Inominados e Não Se Aplica (Secretaria da Turma Recursal).

A variável Entrância separa Comarcas de Entrância Final das de Entrância Inicial, assim divididas:

- Entrância Final: Comarcas de Macapá e Santana;
- Entrância Inicial: Comarcas de Amapá, Calçoene, Ferreira Gomes, Laranjal do Jari, Mazagão, Pedra Branca do Amapari, Porto Grande, Tartarugalzinho e Vitória do Jari.

Para quantificar essas variáveis categorias, o principal método que temos é transformar esses dados categóricos em variáveis binárias, também chamadas de Dummy, assumindo valor 0 ou 1 para indicar a ausência ou presença de uma categoria (Ozdemir & Susarla, 2018).

Os dados obtidos foram retirados da Base Nacional de Dados do Poder Judiciário (DATAJUD), e referem-se ao ano de 2021.

Por se tratar de um número grande de variáveis contínuas, 13 no total, optou-se por reduzir a dimensão desses dados com o uso de técnica estatística de Análise de Componentes Principais, mais conhecida por PCA, utilizada para resumir um número grande de variáveis em outro de dimensão menor.

Os componentes obtidos servirão como passo intermediário para a construção do algoritmo de agrupamento, que é o objetivo principal deste estudo.

2.1 Análise de Componentes Principais

A Análise de Componentes Principais (ACP) é uma técnica estatística multivariada que tem como objetivo reduzir um conjunto grande de variáveis em outro menor, de forma que concentre a maior parte da

informação nesse conjunto menor. Essa técnica permite, ainda, estudar a correlação de covariáveis das variáveis, identificando campos latentes ocultos nos dados, além de determinar índices e produzir escores.
Essas novas variáveis, chamadas de componentes principais, são resultantes de combinações lineares das variáveis originais, contendo a mesma informação das primeiras.

A vantagem desse método está na obtenção de combinações lineares usando um novo sistema de coordenadas, obtido por meio da rotação do sistema original, produzindo componentes ortogonais, ou seja, não correlacionados entre si.

Segundo Johnson & Wichern (2007), considere um vetor com p variáveis aleatórias denominadas $X' = \{X_1, X_2, \dots, X_p\}$ com matriz de covariância dada por $\sum$ e matriz de correlação dada por ρ, com autovalores $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$. As equações são assim representadas:

$$\begin{aligned} Y_1 = a_1'X = a_{11}X_1 + a_{12}X_2 + \cdots a_{1p}X_p \\ Y_2 = a_2'X = a_{21}X_1 + a_{22}X_2 + \cdots a_{2p}X_p \\ \vdots \\ Y_p = a_p'X = a_{p1}X_1 + a_{p2}X_2 + \cdots a_{pp}X_p \end{aligned}$$

As matrizes de variância e covariância serão definidas por:

$$Var(Y_i) = a_i' \sum a_i, \quad com\ i = 1,2, \dots, p$$
$$Cov(Y_i, Y_k) = a_i' \sum a_k, \quad com\ i, k = 1,2, \dots, p$$

Dessa forma, os componentes principais são as combinações lineares não correlacionadas e que tenham maior variância possível.

A ACP aplica a Decomposição de Valores Singulares (SVD) para decompor a matriz de covariâncias em autovalores e autovetores.

Os autovetores ou cargas, a_p', nos darão os pesos para as combinações lineares, que irão gerar os componentes principais $Y's$, ortogonais entre si. Já os autovalores, de acordo com Mingot (2005), referem-se a variância de cada componente principal na explicação da variação total dos dados, sendo proporcionais a quantidade de informação retida em $X_1, X_2, \dots, X_p$.

O primeiro autovalor é sempre o maior, e os demais vão decaindo, indicando que explicam cada vez menos a variabilidade total. Essa análise é interessante para avaliar até que ponto é interessante manter cada componente. Para Johnson & Wichern (2007), não há uma definição para determinar o número ótimo de componentes. Para eles, podemos considerar na análise a quantidade total variância explicada, os tamanhos relativos dos autovalores e as interpretações de cada componente.

Considerando que para fins deste estudo temos variáveis que possuem diferentes escalas, optou-se pela padronização das variáveis, deixando-as com média 0 e desvio padrão igual a 1, evitando assim distorções nos resultados apresentados.

Ao final dessa análise serão guardados os scores para serem usados como novas variáveis na análise de agrupamento.

2.2 Análise de Agrupamento (Cluster)

A análise de agrupamentos ou cluster é um conjunto de técnicas que tem como objetivo criar grupos de instâncias que são semelhantes entre si, com base nas características apresentadas, de maneira que as instâncias do mesmo grupo, ou também chamada de cluster, sejam semelhantes entre si, e que instâncias de grupos separados sejam dissimilares.

De acordo com Hair et al. (2009), para conseguir agrupar os dados, são necessárias três questões básica: Como medir a similaridade? Como formar os agrupamentos? Quantos grupos serão formados?

O primeiro passo para elaborar uma análise de agrupamento envolve definir uma medida de distância (dissimilaridade) para variáveis numéricas ou similaridade (semelhança) para variáveis binárias, que servirá de base para atribuir cada observação a um determinado grupo (Favero & Belfiore, 2017).

Para definir qual distância usar, é importante que estejam bem definidas a natureza das variáveis (discretas, contínuas, binárias), as escalas de medição (nominal, ordinal, intervalo, razão) e conhecimento do assunto (Johnson & Wichern, 2007).

Para este estudo, temos variáveis de tipo mistas, contínuas (resultantes dos componentes principais) e as variáveis dummy que serão agregadas nesta etapa da análise.

Segundo Hunt & Jorgensen (2011), caso os dados sejam de tipo mistos, é necessário utilizar uma medida de distância híbrida, que considera em seu cálculo tanto as variáveis contínuas e binárias.

Para a construção da matriz de distâncias, utilizaremos neste estudo a distância de Gower. Esta medida permite analisar dados de tipos mistos, por meio de um algoritmo que estima a similaridade entre dois objetos (Moura & et al., 2010). Sua expressão é dada por:

$$S_{ij} = \frac{\sum_{k=1}^{p} W_{ijk} \,.\, S_{ijk}}{\sum_{k=1}^{p} W_{ijk}}$$

Onde,

Sijk = contribuição da variável k na similaridade entre os indivíduos i e j, com valores entre 0 e 1.
K = o número de variáveis (k = 1, 2, …, p);
Wijk = peso dado à comparação ijk, atribuindo valor 1 para comparações válidas e valor 0 para comparações inválidas;

2.2.1 Algoritmos para análise de cluster

Dispomos na literatura de dois tipos de algoritmos de agrupamento: hierárquicos e não hierárquicos.

Os algoritmos hierárquicos caracterizam-se por manter uma estrutura de aglomerações para a formação dos agrupamentos (dendrograma), enquanto os não hierárquicos utilizam algoritmos para maximizar a homogeneidade dentro de cada agrupamento, sem que haja um processo hierárquico para tal **(Favero & Belfiore, 2017)**.

Dentre os algoritmos não hierárquicos mais conhecidos destaca-se o K-means, que agrupa os dados baseando-se na média dos valores pertencentes a cada cluster. É aplicado quando as variáveis sob análise são quantitativas e a dissimilaridade é baseada na distância Euclideana. Como este estudo propõe o uso da distância de Gower, este algoritmo se apresenta limitado, e por esta razão não será apresentado com mais detalhes.

2.2.2 Algoritmos hierárquicos

Os algoritmos hierárquicos se dividem em aglomerativos ou divisivos, dependendo do modo como é iniciado o processo. Nos métodos hierárquicos aglomerativos, cada objeto representa um cluster, e assim o número de cluster inicia igual ao de objetos. Já nos métodos divisivos, partimos de um único cluster que contém toda a amostra, que é sucessivamente dividido em subgrupos **(Johnson & Wichern, 2007)**.

Para a construção do algoritmo aglomerativo, o primeiro passo consiste no cálculo da matriz de distâncias, em seguida na identificação de dois cluster mais semelhantes entre si, de modo que esses dois passam a ser um único cluster. No passo seguinte calcula-se novamente as distâncias, considerando os cluster remanescentes. E repete os cálculos de forma sucessiva até encontrar um único cluster.

Ao longo dessas etapas de agrupar pares de cluster, é necessário definir o método que será utilizado para calcular dissimilaridade entre cluster, ou seja, a distância intra-cluster. Alguns deles serão apresentados na sequência:

a) Single linkage - É a menor distância entre as observações A e B.

$$d(A,B) = \min\{d(x_i, x_{i'}\}, para\ x_i \in A, x_{i'} \in B$$

b) Complete linkage - É a maior distância entre as observações A e B.

$$d(A,B) = \max\{d(x_i, x_{i'}\}, para\ x_i \in A, x_{i'} \in B$$

c) Average linkage – É a distância média entre as observações A e B.

$$d(A,B) = \frac{1}{n_A n_B} \sum_{i=1}^{n_A} \sum_{i=1}^{n_B} d(x_i, x_{i'})$$

d) Centroide - É a distância euclideana entre as observações A e B.

$$d(A,B) = d(\bar{x}_A, \bar{x}_B\}$$

e) Ward – É a soma das distâncias ao quadrado entre as observações A e B.

$$d(A,B) = \frac{n_A . n_B}{n_A + n_B} \sum_{i=1}^{n_A} \sum_{i=1}^{n_B} d(x_i - x_{i'})^2$$

2.2.3 Medidas de qualidade de agrupamento

Uma das medidas mais utilizados na literatura para medir a qualidade do agrupamento, é soma de quadrados intra-cluster, definida como a diferença de cada observação em relação ao centro do cluster. Kassambara (2017) define a variação total dentro do cluster da seguinte forma:

$$SQE_A = \sum_{i=1}^{n_A} (\bar{x}_i - \bar{x}_A)^2$$

Portanto, quanto menor o valor da soma de quadrados intra-cluster, maior é a qualidade do agrupamento. Ressaltando que essa medida é parte importante do que subsidia a escolha do número ótimo de cluster.

2.2.4 Número ótimo de cluster

Para Kassambara (2017), determinar o número ótimo de clusters é fundamental para o problema, mas essa análise é algo subjetivo e depende da distância e de método aglomerativo escolhidos.

Existem diferentes métodos para determinar o número ideal de cluster para agrupamentos hierárquicos. Neste estudo, vamos tratar de 2 métodos, são eles:

a) Método Elbow

O método Elbow é conhecido como método do cotovelo. Basicamente o que o método faz é analisar soma de quadrados intra-cluster em função do número de clusters. Essa relação é plotada em um gráfico,

chamado de Elbow Plot, onde podemos localizar o "cotovelo", que corresponde ao ponto onde corre uma mudança acentuada da curva. Esse ponto no gráfico é geralmente considerado como um indicador para o número apropriado de clusters (Kassambara, 2017).

b) Método da Silhueta Média

O método da Silhueta é utilizado para medir a qualidade de um agrupamento. Consiste no cálculo e representação gráfica de uma medida, que determina quão bem cada objeto está alocado ao respectivo cluster, e o quão próximo cada ponto em um cluster está dos pontos nos clusters vizinhos (Kassambara, 2017).

A medida da silhueta é definida considerando o seguinte modelo:

$$s(i) = \frac{b(i) - a(i)}{\max\big(a(i), b(i)\big)}, \ \ para\ i = 1,2, \dots, n.$$

Onde:

- a(i) é distância média de um elemento i em relação a todos os elementos do cluster ao qual ele foi alocado;

b(i) é o menor valor médio das distâncias do elemento i aos elementos de um cluster B, diferente daquele ao qual o elemento i foi alocado;

Esse método calcula a silhueta média das observações para diferentes valores k de cluster, e estima a distância média entre os clusters. O número ótimo de clusters é aquele que maximiza a média em uma faixa de valores possíveis para k (Kassambara, 2017).

Para fins de interpretação, Kassambara (2017) define:

- Observações com um grande s(i) (próximo de 1) estão muito bem agrupadas.
- Um s(i) pequeno (em torno de 0) significa que a observação está entre dois clusters.
- Observações com Si negativo provavelmente são colocadas no cluster errado.

3 Resultados e Discussões

3.1 Análise de Componente Principais

A análise de componentes principais foi realizada utilizando-se o software R versão 4.2.0, através da função **prcomp** do pacote **stats**. Nesta função, foram utilizados os argumentos disponíveis para centralizar e padronizar as variáveis, deixando-as com média 0 e desvio padrão 1.

Os resultados obtidos são apresentados na sequência.

a) Importância dos componentes e variância explicada

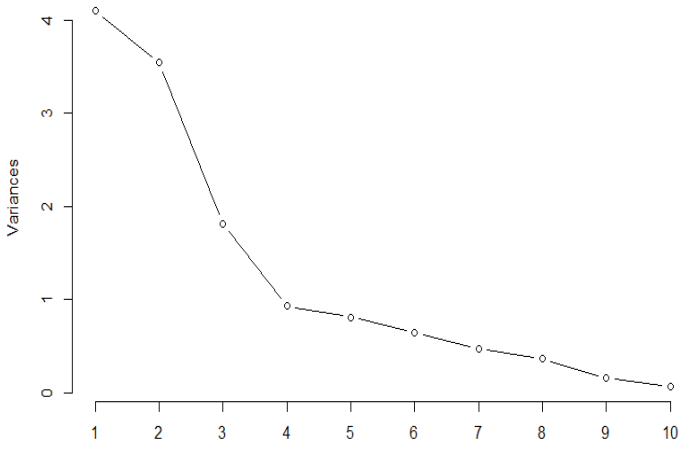
Na Tabela 2 são apresentados os autovalores, em ordem decrescente, definindo a importância das componentes principais.

Tabela 2: Autovalores e Variância Total

Componente Principal	Autovalores	Variância Explicada (%)	Variância Explicada Acumulada (%)
1	4,1006	0,3154	0,3154
2	3,5442	0,2726	0,5880
3	1,8131	0,1395	0,7275
4	0,9285	0,0714	0,7990
5	0,8114	0,0624	0,8614
6	0,6475	0,0498	0,9112
7	0,4773	0,0367	0,9479
8	0,3622	0,0279	0,9757
9	0,1575	0,0121	0,9879
10	0,0711	0,0055	0,9933
11	0,0385	0,0030	0,9963
12	0,0347	0,0027	0.9989
13	0,0137	0,0011	1,0000

Observando o Gráfico 1 do Sreeplot, temos que até o quarto componentes é retida muita informação contida nesses dados, e a partir do quinto componente os autovalores começam a formar quase uma linha reta, com autovalores relativamente pequenos. Portanto, neste trabalho, optou-se por utilizar 4 componentes, que explicam 79,90% para representar o conjunto das 13 variáveis utilizadas nesta análise.

Figura 1 – Gráfico Screeplot



b) Interpretação dos componentes

Na Tabela 3 são apresentados os autovetores, indicando que quanto maior o valor absoluto do

coeficiente, mais importante será a variável correspondente ao calcular o componente.

Tabela 3: Autovalores e Variância Total

Variáveis	PC1	PC2	PC3	PC4
CN	-0,4643	0,0792	-0,0430	0,1445
Sent	-0,4559	0,1087	-0,1236	-0,0238
IPS	-0,4594	-0,1175	-0,0695	0,0533
Tpl	-0,3842	-0,2466	-0,0633	0,1700
Decisões	-0,2611	-0,0428	-0,2980	0,0754
Aud	-0,0947	0,1624	0,5908	-0,0242
AudConc	-0,1483	0,1656	0,5199	-0,3139
Tcl	0,3206	0,0619	-0,3602	0,1075
Iad	-0,0125	-0,4744	0,1308	-0,0940
TpSent	0,0515	-0,4148	0,0605	0,0467
Tbaix	0,0243	-0,4663	0,1300	0,0151
TpCpL	0,0759	-0,4705	0,1331	0,0523
hab.km2	-0,0825	-0,1088	-0,2844	-0,9036

O primeiro componente principal, que tem 31,54% de explicação, tem grandes associações com as variáveis Processos Novos, Processos Julgados, IPSJud, Processos Pendentes líquidos, Decisões e Taxa de Congestionamento Líquido, portanto este componente mede, principalmente, a capacidade das unidades judiciárias em gerenciar os processos. O segundo componente tem grandes associações negativas com o Índice de atendimento à demanda, Tempo médio de julgamento do processo, Tempo médio de baixa do processo, e Tempo médio do processo pendente. Este segundo, portanto, mede principalmente a celeridade na prestação jurisdicional. O terceiro componente tem grandes associações positivas com o número de Audiência e as Audiência de Conciliação. Por fim, o quarto componente associação a Variação da densidade demográfica.

3.2 Algoritmos de agrupamento

Nesta fase da análise, além dos quatro componentes principais obtidos na primeira etapa, foram agregadas ao modelo as variáveis categóricas nominais, Classificação da unidade judiciária (Classe) e Entrância.

Para estruturar a base de dados, as variáveis categóricas foram transformadas em dummies, gerando assim 20 novas variáveis, com valores das observações 0 ou 1.

O primeiro passo para a construção do algoritmo de cluster, é calcular a matriz de distâncias. Com essa medida é possível analisar o quão distantes estão as observações em um conjunto de dados. Para isso, foi utilizada a função **dayse** do pacote **cluster**, do software R. A grande vantagem dessa função é conseguir trabalhar com variáveis de diferentes escalas.

A Tabela 4 mostra a matriz de distância de Gower para as cinco primeiras unidades judiciárias, de um total de 59. Nela podemos observar que as cinco primeiras unidades judiciárias da tabela de dados são parecidas, pois as distâncias são pequenas. E isso reflete a realidade, pois são unidades que tem competências predominantemente Cíveis e de Fazenda Pública, pertencentes a mesma entrância e localidade, além de terem seus dados de gestão processual produtividade bem parecidos.

Tabela 4: Distâncias da Gower

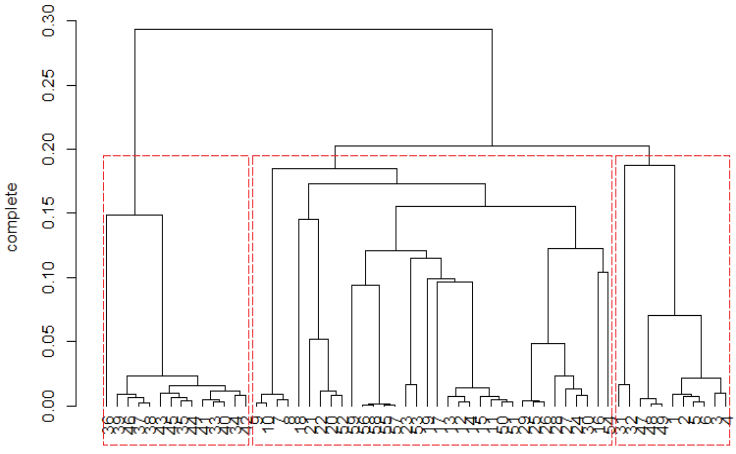
	1	2	3	4	5
1	0,0000				
2	0,0090	0,0000			
3	0,0154	0,0074	0,0000		
4	0,0210	0,0143	0,0095	0,0000	
5	0,0083	0,0051	0,0108	0,0193	0,0000

O próximo passo foi calcular o dendrograma, uma ferramenta gráfica que mostra como os clusters são criados em cada passo, e qual o nível de similaridade dos grupos formados.

Para a construção do dendrograma foi necessário escolher o método para a distância intra-cluster. E após testar diferentes distâncias, foi adotada a ***Complete***, pois foi a que apresentou o dendrograma mais equilibrado em relação aos demais, embora o agrupamento gerado por essas distâncias tenham sido praticamente os mesmos.

Ao analisar o dendrograma da Figura 2, podemos verificar que se traçarmos uma linha de corte em torno da altura 0,20, criaríamos 3 clusters. Esses três grupos seriam assim classificados: 1) unidades de entrância inicial; 2) unidades de entrância final com competências cíveis e/ou fazenda pública; 3) demais unidades de entrância final.

Figura 2 – Dendrograma com 2 grupos destacados



Esses grupos formados nesses não se mostraram

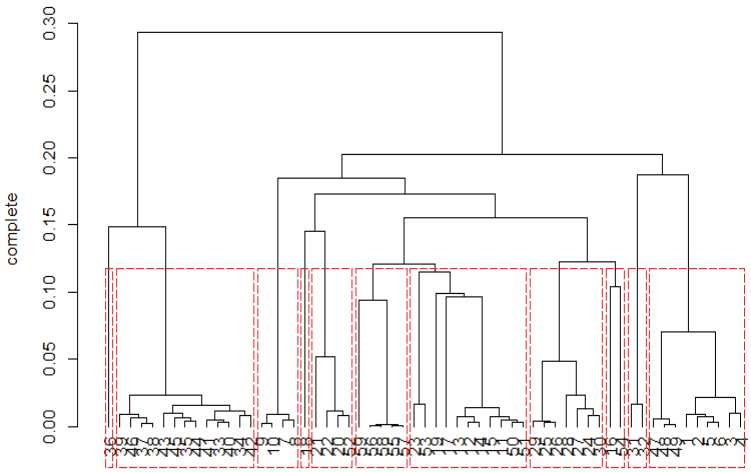
razoável, pois do ponto de vista do negócio, classificar unidades judiciárias de entrância final que apresentam características bastante diferentes no mesmo grupo, não seria aconselhável.

Após vários testes, concluímos que o mais coerente é traçarmos uma linha de corte em torno da altura 0,12. Dessa forma, teremos a formação de 11 grupos, classificados conforme a Tabela 5, a apresentados graficamente na Figura 3.

Tabela 5: Formação dos 11 grupos e descrição da competência

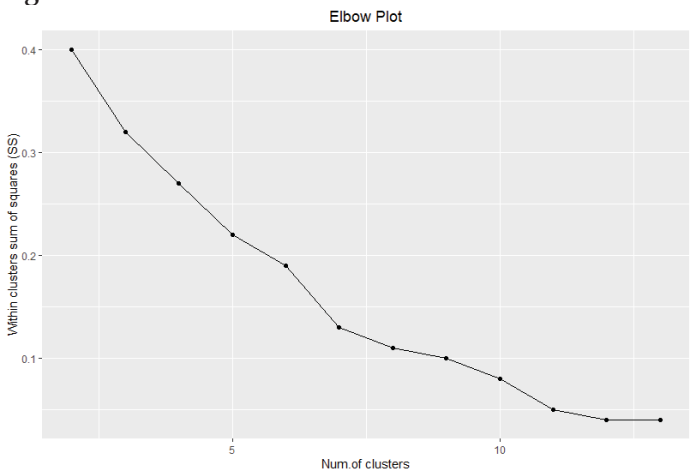
Grupos	Entrância	Descrição da Competência
Grupo 1	Final	Cíveis, Cíveis e Fazenda Pública
Grupo 2	Final	Família, Órfãos e Sucessões
Grupo 3	Final	Criminal, Tribunal do Júri, Execução de Penas e Medidas Alternativas, Violência Doméstica e Familiar Contra a Mulher
Grupo 4	Final	Juizado Especial Criminal, Juizado Especial Cível e Criminal
Grupo 5	Final	Execução Penal
Grupo 6	Final	Infância e Juventude
Grupo 7	Final	Juizado Especial Cível
Grupo 8	Final	Juizado Especial da Fazenda Pública
Grupo 9	Final	Recursos Inominados e Não se Aplica (Secretaria da Turma Recursal)
Grupo 10	Inicial	Juízo Único
Grupo 11	Inicial	Juizado Especial Cível, Criminal e da Fazenda Pública

Figura 3 – Dendrograma com 11 grupos destacados



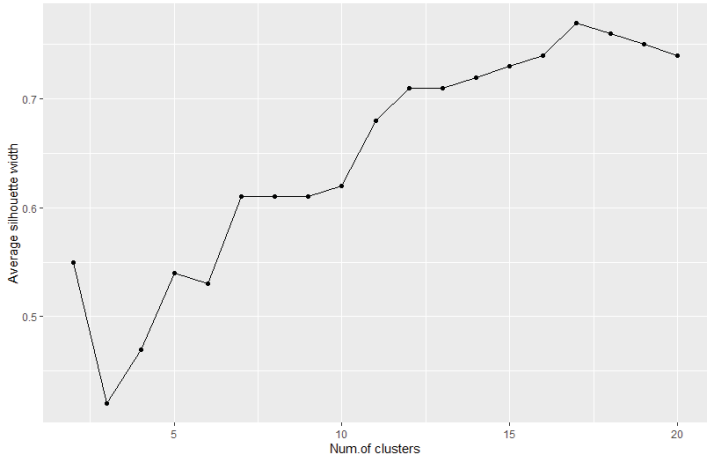
Para verificarmos o número ótimo de cluster, utilizaremos do método Elbow, através da análise do gráfico Elbow Plot. Observando a Figura 4, é possível verificar que o número ótimo de clusters sugerido para agrupar as unidades judiciárias é 11, pois a partir desse ponto há um ganho baixo em adicionar de um novo cluster.

Figura 4 – Elbow Plot



Outro método utilizado para verificar o número ótimo de cluster é o da Silhueta. De acordo com este método, representado graficamente na onde na Figura 5, o número ótimo de cluster é 17, que representa o ponto mais alto no gráfico.

Figura 5 – Gráfico da Silhueta



Considerando a classificação do 17 clusters, de acordo com o método da Silhouete, teremos na Tabela 6 a seguinte disposição das unidades judiciárias nos grupos:

Tabela 6: Formação dos 17 grupos e descrição da competência

Grupos	Entrância	Descrição da Competência
Grupo 1	Final	Cíveis e Fazenda Pública
Grupo 2	Final	Família, Órfãos e Sucessões
Grupo 3	Final	Criminal
Grupo 4	Final	Juizado Especial Criminal
Grupo 5	Final	Tribunal do Júri
Grupo 6	Final	Execução Penal

Grupo 7	Final	Execução de Penas e Medidas Alternativas
Grupo 8	Final	Infância e Juventude
Grupo 9	Final	Violência Doméstica e Familiar Contra a Mulher
Grupo 10	Final	Juizado Especial Cível
Grupo 11	Final	Juizado Especial da Fazenda Pública
Grupo 12	Final	Cíveis
Grupo 13	Inicial	Juizado Especial Cível, Criminal e da Fazenda Pública
Grupo 14	Final	Recursos Inominados
Grupo 15	Final	Não se aplica (Secretaria da Turma Recursal)
Grupo 16	Inicial	Juízo Único
Grupo 17	Inicial	Juizado Especial Cível e Criminal

O que o método da Silhueta propôs foi a classificação das unidades judiciárias em cluster ainda mais distintos, considerando principalmente a classificação das competências predominantes.

4 Conclusões

Considerando o cenário da Justiça Nacional, verifica-se que diferentes tribunais adotam critérios subjetivos para agrupar unidades judiciárias. Neste sentido, a proposta deste estudo foi trazer um modelo estatístico, com uso de um algoritmo de agrupamento, que determinasse unidades semelhantes por critérios objetivos, utilizando como exemplo dados do Tribunal de Justiça do Amapá. Os resultados apresentados apontam pela eficiência dos métodos utilizados, uma vez que conseguiu agrupar e diferenciar, a partir das informações de entrada, unidades similares e dissimilares entre si, respectivamente.

Agradecimentos

Agradeço primeiramente a coordenação do Curso pelo aceite da minha candidatura. Grata à oportunidade de ter aprendido com professores tão sábios e dedicados na formação de Cientistas de Dados.
Agradeço ao meu orientador, Professor Dr. Fernando Mayer, pelo aceite do convite, pela paciência e orientação.
Agradeço as minhas colegas do Tribunal de Justiça do Amapá, Tayanny Negrão e Rubia Balieiro, pela força e ajuda na realização desse trabalho.
Não deixaria de agradecer a minha família que sempre está do meu lado, me apoiando na realização dos meus projetos.

Referências

Tribunal de Justiça do Estado do Amapá. Comarcas. Fonte: https://www.tjap.jus.br/portal/home/comarcas.html

Conselho Nacional de Justiça. Resolução 76/2009. Fonte: https://atos.cnj.jus.br/atos/detalhar/110

DATAJUD. Base Nacional de Dados do Poder Judiciário. Fonte: Base Nacional de Dados do Poder Judiciário

Favero, L. P., & Belfiore, P. (2017). *Manual de Análise de Dados - Estatística e Modelagem Multivariada com Excel®, SPSS® e Stata®.* Rio de Janeiro: Elsevier.

Hair, J. F., Black, W., Babin, B., Anderson, R., & Tatham, R. (2009). *Análise Multivariada de Dados* (6ª Edição ed.). Bookman.

Hunt, L., & Jorgensen, M. (2011). *Clustering mixed data.* Wiley Interdisciplinary Reviews: Data.

Johnson, R. A., & Wichern, D. (2007). *Applied Multivariate Statistical Analysis* (6th Edition ed.). Pearson Prentice Hall, Upper Saddle River.

Kassambara, A. (2017). *Practical Guide To Cluster Analysis in R.* Sthda.

Mingot, S. A. (2005). *Análise de dados através de métodos de estatística multivariada.* UFMG.

Moura, M. d., & et al. (abr de 2010). Algoritmo de Gower na estimativa da divergência genética em germoplasma de pimenta. *Horticultura Brasileira*, 155-161.

Ozdemir, S., & Susarla, D. (2018). *Feature Engineering Made Easy.* Packt.