UFRJ

UNIVERSIDADE FEDERAL DO RIO DE JANEIRO INSTITUTO DE MATEMÁTICA
PÓS-GRADUAÇÃO EM ESPECIALIZAÇÃO EM CIÊNCIA DE DADOS

THAMIRYS DO POÇO CHAVES DIAS
PASCHOAL FALCONI JÚNIOR

Aprendizado de máquina aplicado a dispersão geográfica de carteira de seguro agrícola para a cultura de soja das seguradoras

TRABALHO DE CONCLUSÃO DE CURSO

Rio de Janeiro
2022

THAMIRYS DO POÇO CHAVES DIAS
PASCHOAL FALCONI JÚNIOR

Aprendizado de máquina aplicado a dispersão geográfica de carteira de seguro agrícola para a cultura de soja das seguradoras

Trabalho de Conclusão de Curso apresentado a Pós-graduação em Especialização em Ciência de Dados do Instituto de Matemática da Universidade Federal do Rio de Janeiro como parte dos requisitos necessários à formação de Especialista em Ciência de Dados.

Orientador: João Antonio Recio da Paixão, DSc

Rio de Janeiro
2022

Dedicamos este trabalho à Deus e a todos os familiares e amigos que estiveram presentes na jornada da nossa formação nesta especialização.

Agradecimentos

Gostaríamos de registrar os nossos mais sinceros agradecimentos:

- A Deus que nos deu força para concluir mais etapa em nossas vidas;
- Aos nossos amados familiares pelo apoio incondicional nessa jornada que se passou no meio de uma pandemia e todas as adversidades que surgiram no meio do percurso;
- Aos queridos amigos que participaram de forma direta e indireta de nossa formação;
- Aos colegas de turma, que mesmo de forma virtual criamos laços e parcerias;
- Aos nossos colegas de trabalho pelo incentivo na busca desse desenvolvimento profissional.

Resumo

O objetivo deste trabalho de conclusão de curso é apresentar uma metodologia, através da utilização de aprendizado de máquina, de gerenciamento de carteira de seguro agrícola das seguradoras com foco na dispersão geográfica do risco, visando equilibrar o resultado das companhias.

O seguro agrícola tem como um dos objetivos proteger o produtor rural contra perdas decorrentes de fenômenos climáticos adversos que afetem a lavoura e que acarrete queda de produção. Em resumo, proporciona a recuperação do valor investido na lavoura, mantendo a cadeia do agronegócio saudável e sendo uma importante ferramenta de política agrícola. Uma vez que um problema climático irá possivelmente acometer mais de uma lavoura exposta naquela região se caracterizando como um evento catástrofe. Como exemplo, o déficit hídrico ocorrido na safra de verão 2021/2022 na Região Sul e no Mato Grosso do Sul, influenciado pelo fenômeno La Niña, resultando em redução de produtividade de soja nesses locais.

Vale ressaltar, que este estudo não abordará as particularidades dos produtos de seguro oferecidos no mercado, mas sim, vislumbra realizar uma análise de agrupamento dos municípios com similaridade de potencial produtividade, representatividade em termos de área plantada, histórico de sinistralidade do Programa de Subvenção ao Prêmio o Seguro Rural (PSR) e histórico de precipitação por estado para a cultura de soja, possibilitando auxiliar na tomada de decisão em termos de diversificação geográfica da carteira de seguro agrícola.

Palavras-chave: Dispersão geográfica. Similaridade. Seguro agrícola. Aprendizado de máquina. Análise de agrupamento.

Abstract

The objective of this paper is to present a methodology, by the utilization of the machine learning, for managing the insurance corp portfolio of insurers with a focus on the geographic dispersion of risk, aiming to balance the results of the companies.

One of the objectives of insurance corp is to protect the rural producer against losses resulting from adverse weather phenomena that affect the crop and lead to a drop in production. In short, it provides the recovery of the amount invested in farming, keeping the agribusiness chain healthy and being an important agricultural policy tool. Since a climate problem will possibly affect more than one crop exposed in that region, characterizing it as a catastrophe event. As an example, the rain deficit that occurred in the 2021/2022 summer harvest in the South Region and Mato Grosso do Sul, influenced by the La Niña phenomenon, resulting in a reduction in soybean productivity in these locations.

It is worth mentioning that this study will not address the particularities of the insurance products offered in the market, but rather, it envisages performing a grouping analysis of municipalities with similarity of potential productivity, representativeness in terms of crop area, history of claims of the Subsidy Program to the Rural Insurance Premium (PSR) and rainfall history per state for soybeans, making it possible to assist in decision making in terms of geographic diversification of the agricultural insurance portfolio.

Keywords: *Geographic dispersion. Similarity. Crop insurance. Machine learning. Cluster analysis.*

Lista de figuras

Sumário

1 Introdução

O agronegócio tem um importante papel na economia do país, reflexo disso é a sua representatividade no PIB (Produto Interno Bruto) Brasileiro que em 2020 ficou em torno de 27%. Essa cadeia engloba os três setores da economia:

- Primário: Produtores ruais, independente do seu tamanho;
- Secundário: Agroindústrias e fornecedores de insumos;
- Terciário: Engloba a cadeia de distribuição e exportação.

Existem ainda outros setores que se relacionam mesmo que de forma indireta. Por exemplo, Bancos, farmacêuticas, entre outras.

A principal cultura produzida no Brasil é a soja, o que nos levou a ser o maior produtor deste grão na safra de 2020/2021, com produção de 135,409 milhões de toneladas frente a produção de 112,549 milhões de toneladas do EUA.

Apesar do crescente avanço tecnológico em campo e do aprimoramento das técnicas de manejo, a agricultura está sujeita a fenômenos climáticos adversos que podem acarretar perdas significativas na produção. O que a classifica como uma atividade de alto risco, por se tratar de fenômenos que não são possíveis de serem controlados pelo ser humano. Outra característica desta cadeia e consequência de problemas climáticos na agricultura é que no geral a ocorrência desses eventos afeta várias lavouras que estão naquela região. Por exemplo, na safra de verão 2021/2022, a região Sul e o estado do Mato Grosso do Sul, enfrentou um déficit hídrico severo. Em termos monetários, os valores pagos pelas seguradoras aos produtores rurais no primeiro trimestre de 2022 foi de R$ 5,8 bilhões, segundo o Ministério da Agricultura, Pecuária e Abastecimento (MAPA) (Fonte: https://cnseg.org.br/noticias/governo-divulga-portaria-sobre-seguro-rural.html). Ou seja, se trata de um ramo suscetível a catástrofes naturais.

Neste caso, o seguro agrícola é um instrumento de transferência de risco que tem como papel minimizar as perdas oriundas dessas intempéries e possibilitar que o produtor rural recupere o valor investido na lavoura, evitando também renegociação de dívidas, assegurando o incremento de investimento e consequentemente de produção, e reduzindo os impactos a outros setores envolvidos na cadeia do agronegócio.

Uma das políticas públicas do Governo Federal é o Programa de Subvenção de Seguro Rural (PSR), onde ele apoia o produtor rural subsidiando parte do prêmio da apólice, com a finalidade de massificar a utilização do seguro rural; garantir o cumprimento do papel desta ferramenta como mitigador de risco climático; fomentar a utilização de tecnologia e modernização da gestão do campo, uma vez que busca estabilizar a renda do produtor rural. Vale ressaltar que existem critérios para que o produtor seja elegível a esse programa, bem como disponibilização de verba.

Avaliando o cenário das Seguradoras, um dos desafios é através do mutualismo; conceito básico do seguro que visa a formação de um grupo de pessoas expostas a riscos similares para a formação de um fundo com o objetivo de arrecadar prêmios e que suportará as indenizações caso ocorram sinistros; e da característica catastrófica do seguro agrícola, realizar a dispersão de risco de forma geográfica. Desta forma, em um ano sob influência de La niña, evento climático que traduz em maior probabilidade de ocorrência de secas no sul do país, haja uma compensação com os riscos expostos em outras regiões e que não tem o regime das chuvas impactado por este fenômeno.

2 Objetivo

O objetivo desse trabalho é avaliar uma metodologia aplicando conceitos de aprendizado de máquina para dispersão geográfica da carteira de seguro agrícola de uma seguradora, para a cultura de soja, buscando um gerenciamento de risco que equilibre o resultado financeiro da companhia.

O método avaliado foi o de clusterização, análise não supervisionada, K-Means, considerando as seguintes variáveis para cada município/estado a fim de agrupar por estado municípios que possuem características similares.

- Rendimento médio de produção de soja e quilogramas por hectare de 2010 a 2020 disponibilizado pelo Instituto Brasileiro de Geografia e Estatística (IBGE);

- Área plantada de soja em hectare em 2020 disponibilizado pelo Instituto Brasileiro de Geografia e Estatística (IBGE);

- Sinistralidade da cultura de soja do Programa de Subvenção de Seguro Rural de 2010 a 2020 disponibilizado pelo Ministério da Agricultura, Pecuária e Abastecimento (MAPA) e a área contratada em 2021;

    - Sinistralidade: É a razão entre indenizações e prêmios;

    - Prêmio: Valor pago pelo segurado à seguradora para ter direito a cobertura do seguro;

    - Indenização: Pagamento feito pelo segurado ao segurado quando o seguro é acionado.

  Caso esse resultado seja superior a 100%, indica que a operação está tendo prejuízo, uma vez que a quantidade de indenizações pagas está sendo maior que o valor arrecadado com o prêmio.

  A inclusão desse parâmetro tem como objetivo captar o risco inerente das operações de seguro em cada município e que está correlacionado com os problemas climáticos que afetaram aquela região.

- Dados reanálises de precipitação por município disponibilizado pelo CHIRPS: O regime de chuvas é um fator fundamental para a produtividade agrícola. As safras podem ser seriamente afetadas em função de índices pluviométricos muito abaixo da média. Desta forma, se torna preponderante, em uma análise que tem como foco o comportamento da produtividade, a utilização de dados pluviométricos. Uma vez que a disponibilidade de estações pluviométricas dos órgãos oficiais é limitada a poucos municípios do território, a estratégia deste estudo foi utilizar dados de reanálise por satélite do “CHIRPS” (Climate Hazards Group InfraRed Precipitation). Reanálise meteorológica é um conjunto de dados obtidos a partir de modelos de circulação global com dados medidos, agrupando os dados disponíveis em um contexto físico.

  Entre os dados de reanálise, O CHIRPS foi desenvolvido pelo UNITED States Geological Survey (USGS) e pelo Climate Hazards GROUP at the University of California, santa Barbara (UCSB) e é uma base de dados de precipitação, com dados desde janeiro de 1981, composta por diferentes fontes de informação, tais como de satélites com espectroscopia de infravermelho termal (Thermal Infrared, TIR), geoestacionárias quase globais da National Oceanic and Atmospheric Administration (NOAA); Centro de Previsão Climática

(CPC); National Climatic Data Center Climáticos (NCDC); Coupled Forecast System da NOAA, versão 2 (CFSv2) e dados observacionais de estações meteorológicas (Funk et al., 2015).

O produto CHIRPS apresenta como características resolução espacial de 0,05º, o que corresponde a aproximadamente 5 km, com uma cobertura geográfica de 50º S a 50º N e conta com uma base de dados de 1981 até os dias atuais (com dados diários, pentadados e mensais). Os dados do CHIRPS são disponibilizados de forma gratuita no site ftp://ftp.chg.ucsb.edu/pub/org/chg/products/CHIR PS-2.0/, nos formatos NetCDF, GeoTiff e Esri BIL.

Fontes: Análise do Desempenho da Estimativa de Precipitação do Produto CHIRPS para Sub-Bacia do Rio Apeú, Castanhal-PA Emerson Renato Maciel da Silva1; Ivan Carlos da Costa Barbosa2; Helder José Farias da Silva3; Luiz Gonzaga Silva Costa4; Edson José Paulino da Rocha5.

3 Metodologia

Aprendizado não-supervisionado:

No aprendizado não supervisionado, os dados não são rotulados. Ou seja, não há uma variável resposta (ou “target), apenas um conjunto de variáveis explicativas. O objetivo para cada elemento não será, então, obter uma resposta categórica ou numérica específica e sim encontrar semelhanças entre os indivíduos do conjunto de dados.

Desta forma, algoritmos de aprendizado não-supervisionado é um tipo de algoritmo que aprende ou constrói padrões a partir de dados não marcados. O que se espera é que através de características comuns entre os indivíduos, se elabore uma representação interna da população, em contraste com o aprendizado supervisionado, onde os dados são marcados por um especialista, por exemplo.

Assim, quando se deseja entender os dados a partir de suas próprias características a partir daí deduzir modelos ou situações que existem para o problema, o algoritmo precisa encontrar estruturas e padrões, e então classificar os novos resultados a partir da similaridade das suas características, com os grupos descobertos no treinamento.

Modelos de aprendizagem não supervisionada são muito úteis para guiar o raciocínio no processo de exploração de dados para análises futuras.

Alguns dos principais tipos de algoritmos (e seus respectivos representantes) de Aprendizado Não-Supervisionado são os seguintes:

- Clustering:
    - K-Means;
    - Clustering Hierárquico;
    - Maximização da Expectativa.
- Visualização e redução de dimensionalidade:
    - Análise de Componentes Principais (PCA);
    - Kernel PCA;
    - Localli-Linear Embedding (LLE);
    - T-distributed Stochastic Neigboor Embedding(t-SNE).
- Aprendizado da regra da associação:
    - Apriori;
    - Eclat;

Em relação ao tema deste projeto, tem-se como finalidade tentar classificar os municípios quanto ao seu de produtividade de soja, média e volatilidade, a sua representatividade em termos de área plantada, a precipitação média acumulada nos meses da safra e seu coeficiente de variação e sinistralidade de seguro agrícola para a cultura de soja. Para isto, o indicado é a utilização de algoritmos de clusterização.

Clusterização:

A clusterização tem o objetivo agrupar os dados de interesse, de tal forma que elementos de um cluster compartilhem um conjunto de propriedades comuns que os diferencie de outros clusters. São os problemas de aprendizagem não-supervisionada mais comuns.

Neste tipo de problema, a ideia principal é que elementos que componham o mesmo grupo devem ser bastante similares, mas bastante diferentes de indivíduos dos outros "clusters". De outra maneira, o agrupamento é realizado com o intuito de maximizar a homogeneidade dentro de cada cluster e maximizar a heterogeneidade entre os "clusters".

O uso de técnicas de clusterização, ao agrupar dados parecidos, descreve de forma mais eficiente e eficaz as características próprias a cada grupo construído pelo método, o que permite uma maior compreensão do conjunto de dados original, além de possibilitar o desenvolvimento de esquemas de classificação para novos dados e descobrir correlações interessantes entre as "features" dos dados que não seriam facilmente visualizadas sem o emprego da técnica.

Há vários tipos de métodos de clusterização, mas os dois mais tradicionais são os Métodos Particionais e os métodos Hierárquicos:

Métodos Hierárquicos: algoritmos de clusterização que organizam um conjunto de dados em uma estrutura hierárquica de acordo com a proximidade entre os indivíduos. Pode ter uma abordagem aglomerativa (bottom-up) ou divisiva (top-down).

Métodos Particionais: Os algoritmos particionais dividem a base de dados em k-grupos. Os indivíduos são divididos entre os k-clusters de acordo com a medida de similaridade escolhida (distância euclidiana, distância Manhatan etc.). Esses algoritmos utilizam métodos iterativos para determinação dos integrantes dos k-clusters.

O K-means é o mais popular e mais simples algoritmo particional. Por sua simplicidade, facilidade de implementação, eficiência e, será a metodologia de clusterização utilizada neste trabalho.

K-Means:

É um método de clusterização que tem como objetivo a partição das n observações de uma população ou amostra em k-clusters no qual cada observação pertence ao cluster de média mais próxima (centroide do cluster).

A estrutura do algoritmo se baseia inicialmente na seleção aleatória de k pontos do conjunto de dados, que correspondem aos k centroides iniciais dos k "clusters" a serem formados. Em seguida, para cada ponto no conjunto de dados, o algoritmo calcula a distância deste ponto a cada um dos centroides selecionados inicialmente e atribui a este ponto o grupo onde a distância seja menor. Desta forma, cada elemento, nesta primeira etapa, pertencerá a um grupo.

Com os grupos formados, os centroides são recalculados pela média dos elementos do grupo e nova realocação dos indivíduos e realizada. O processo se repete até que não haja mais modificação dos centroides ou após um número pré-estabelecido de iterações;

Resumidamente:

- K pontos selecionados aleatoriamente como centroides;

- Para cada indivíduo, calcula-se a distância aos centroides;
- Atribui-se a o ponto o grupo no qual a distância ao centroide seja a menor;
- Com os grupos definidos, recalculam-se os centroides pelas médias das distâncias indivíduos do grupo ao respectivo centroide;
- Repete-se o procedimento até que não haja mais mudanças de centroide.

A figura abaixo ilustra o funcionamento do método para k (número de clusters) igual a 2:

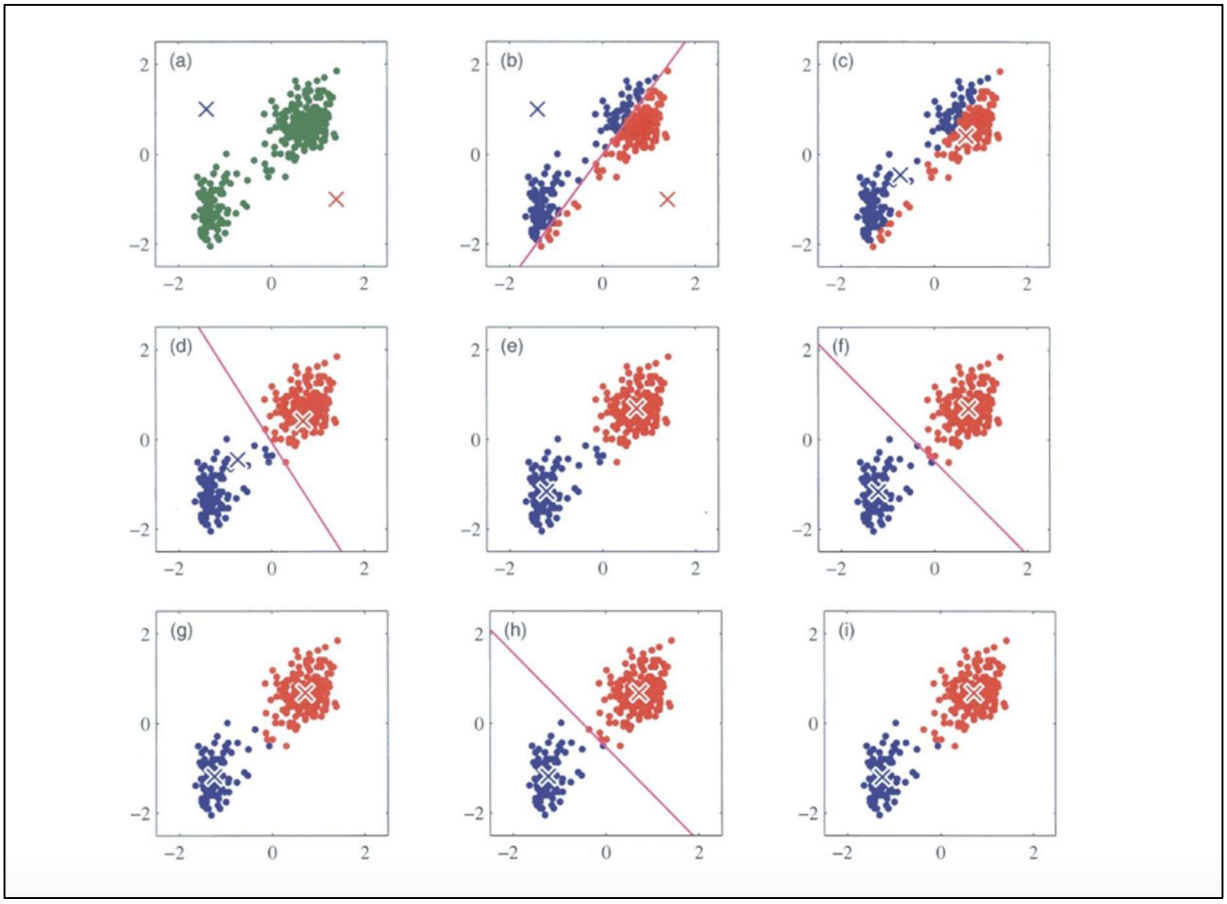

Figura 1: Ilustração do algoritmo K-means de [Bishop 2006]

Na ilustração acima (Bishop, 2006, apud Dendroid), nota-se em (a) que as médias (representadas pelas cruzes) são selecionadas aleatoriamente do conjunto de dados (no caso específico, 2). No exemplo da ilustração, trabalhamos com o espaço bidimensional, mas processo semelhante pode ser estendido para um conjunto de dados n-dimensional. No próximo passo (b), todos os dados são divididos em dois clusters dependendo da proximidade de cada ponto com a média definida inicialmente de forma aleatória. Após a divisão, os centros dos clusters são recalculados para que sejam a média dos pontos atribuídos ao cluster correspondente (c). Os quadros (d)-(h) mostram a melhoria dos centros de cluster ao se repetir iterativamente o processo de (b)-(c). O algoritmo converge para o estado (i), no qual as médias dos clusters obtém uma estabilidade e não mudam mais. Os pontos em (i) então são definidos como representativos e os clusters são selecionados.

Em geral, o método de K-means apresenta bom desempenho quando os grupos são densos, compactos e bem separados uns dos outros, além de ser computacionalmente rápido e de fácil entendimento e implementação. No entanto, há a necessidade de especificar k = número de clusters. Isso pode ser feito de maneira arbitrária, por conhecimento prévio ou por metodologias de validação como análise de Sillhueta (Silhouette).

Análise de Silhueta (Silhouette):

Análise de Silhueta ou Silhouette é um método de interpretação e validação da consistência para análise de “clusters”. A técnica fornece uma representação gráfica de quão boa foi a classificação de cada elemento da amostra/população. Pela análise de Silhueta, pode-se obter-se ainda o indicador de Silhueta para escolha do número ótimo de “clusters”.

O valor de Silhueta é uma medida de quão similar é um indivíduo dentro de seu próprio cluster comparado com os outros clusters. O valor de silhueta varia ente -1 e 1, no qual altos valores indicam um bom ajuste do elemento ao seu próprio cluster. Se muitos indivíduos têm valores de silhueta altos, então a configuração de “clusters” proposta é adequada.

Suponha que os dados são clusterizados via uma técnica qualquer, tal como k-means, em k “clusters”.

Para cada ponto i pertence a CI define-se:

$$a_{(i)}=\frac{1}{|C_I|-1}\sum_{j \in C_I, i \neq j} d(i,j)$$

A média das distâncias entre o indivíduo i e todos os outros pontos no mesmo cluster, no qual $|C_I|$ é o número de pontos pertencentes ao cluster C_I e $d_{(i,j)}$ é a distância entre os pontos i e j no cluster C_I, com j diferente de i, evidentemente.

Pode-se interpretar a(i) como uma medida de quão bem o indivíduo i é relacionado ao seu próprio cluster.

De maneira análoga, define-se a medida de não similaridade do ponto i com o cluster C_J como a média da distância de i para todos os pontos em C_J no qual C_J diferente de C_I. Para cada ponto i pertencente a C_I define-se:

$$b_{(i)}=\underset{J \neq I}{Min}\frac{1}{|C_J|}\sum_{j \in C_I} d(i,j)$$

Como a menor distância média dos do indivíduo i para todos os pontos de qualquer outro cluster do qual i não seja membro. O “cluster” com a maior similaridade é denominado “cluster vizinho”.

Agora então podemos definir o valor de silhueta para um ponto i como:

$s_{(i)} = \frac{b_{(i)} - a_{(i)}}{max\{a_{(i)}, b_{(i)}\}}$, se $|C_I| > 1$ e: $s_{(i)} = 0, se\ |C_I| = 1$

Da fórmula acima, chega-se a:

$$-1 <= s_{(i)} <= 1$$

Observa-se que $a_{(i)}$ não é definido claramente para clusters com número de observações igual a 1, para o qual define-se $s_{(i)} = 0$.

O valor $s_{(i)}$ sobre todos os pontos de um cluster é a medida de como bem agrupados os pontos deste “cluster” estão.

A média de todos os valores $s_{(i)}$ é o coeficiente de silhueta e é utilizado para a definição do número de “clusters”. Assim a escolha do número de cluster será dado por:

$$SC = max_k^{\bar{s}}(k)$$

No qual $\overline{s_{(k)}}$ representa a média dos $s_{(i)}$ sobre todos os indivíduos do conjunto de dados para um número específico de clusters k.

4 Análise descritiva dos dados

A linguagem de programação Python foi utilizada para extrair, tratar e analisar as variáveis consideradas neste trabalho. A seguir tais variáveis foram elencadas, incluídas uma breve descrição a respeito delas, citados alguns tratamentos realizados nos dados e análises exploratórias.

Rendimento médio de produção de soja e quilogramas por hectare de 2010 a 2020:

O rendimento médio de produção por cultura e regiões geográficas do Brasil (Unidades da Federação, Mesorregiões Geográficas, Microrregiões Geográficas e Municípios) é fornecida pelo Instituto Brasileiro de Geografia e Estatística (IBGE) anualmente. Através dessas informações podemos avaliar a produção média de cada local em quilogramas por hectare e identificar padrões como incremento ou queda de produtividade e correlacionar com possível intempéries ocorridas nas safras. Adiante sendo nomeado também como rendimento médio ou produtividade.

Os dados de soja considerados nesse estudo foram extraídos do site Sistema IBGE de Recuperação Automática (SIDRA) na aba Produção Agrícola Municipal (PAM) (https://sidra.ibge.gov.br/pesquisa/pam/tabelas). Para trabalhar o dataframe importado do SIDRA, é necessário ajustar a formatação e tipo das variáveis.

Alguns municípios não possuem a série dos últimos 10 anos (2010 a 2020) de rendimento médio de soja completa, o tratamento de dados ausentes se deu através da substituição deles pela média dos últimos 10 anos de cada município a fim de impactar o menos possível na média total e contemplar locais que possuem histórico de produção agrícola mais recente.

Como segundo tratamento dos dados, para os municípios que de fato continuaram com a média como dado ausente, esses valores foram substituídos por zero, uma vez que não houve produção de soja naquele local. Esta iniciativa foi realizada no âmbito de análise exploratória, para definição dos municípios a serem consideradas na clusterização outros parâmetros foram considerados.

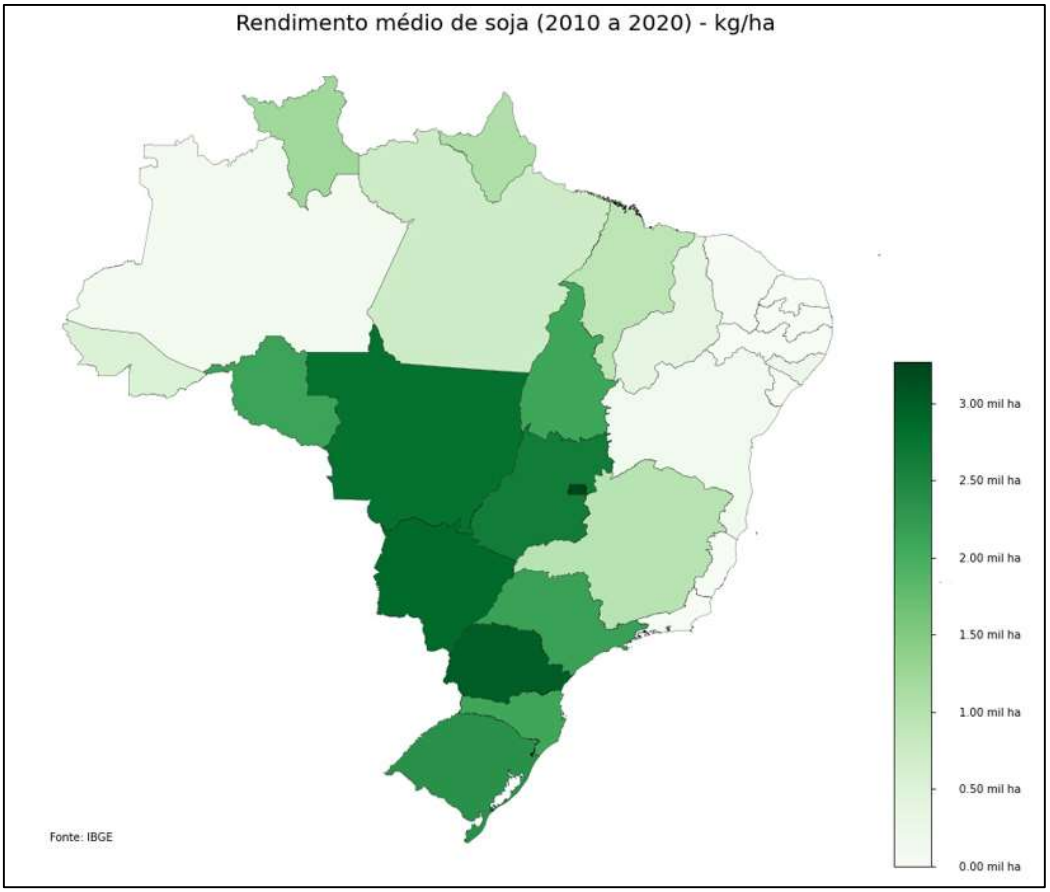

Figura 2: Mapa cloroplético do Brasil de Rendimento médio de soja (2010 a 2020) - kg/ha por UF

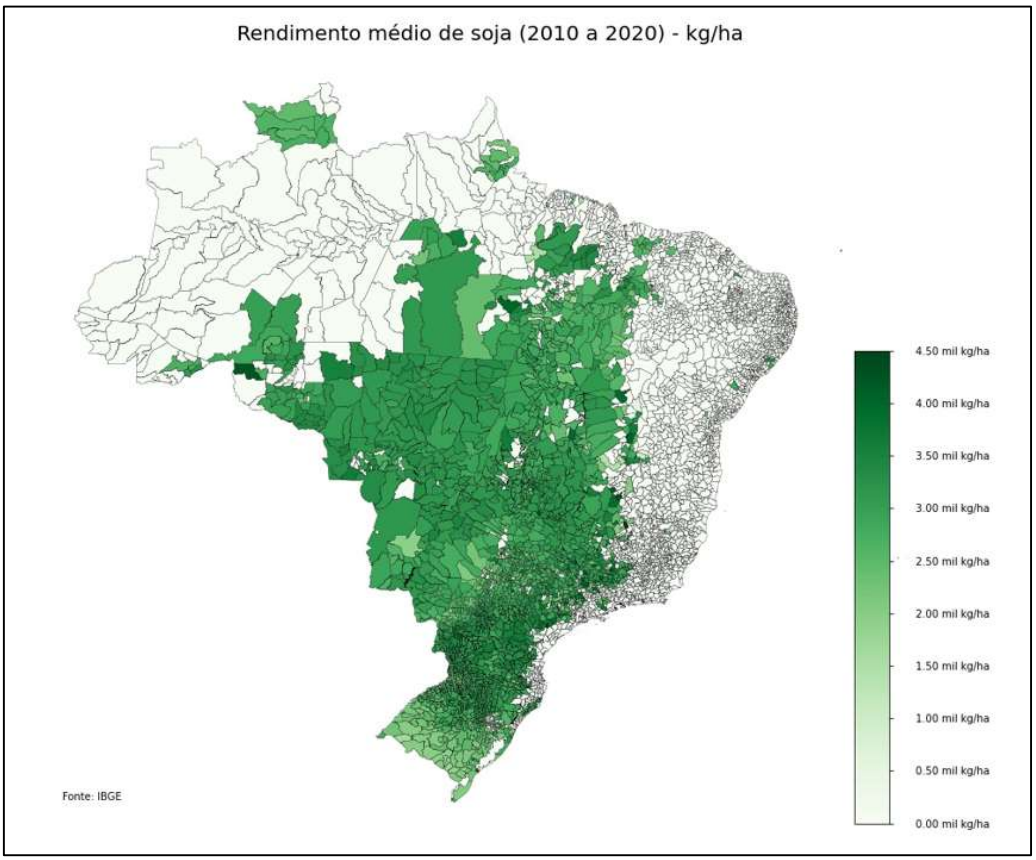

Figura 3: Mapa cloroplético do Brasil de Rendimento médio de soja (2010 a 2020) - kg/há por município

Observa-se, no mapa abaixo algumas regiões com maior volatilidade quanto a rendimento médio de soja. O esperado neste trabalho é que o algoritmo identifique essas nuances e agrupe os municípios em função da sua similaridade, apoiando a tomada de decisão em uma definição de estratégia de dispersão de risco.

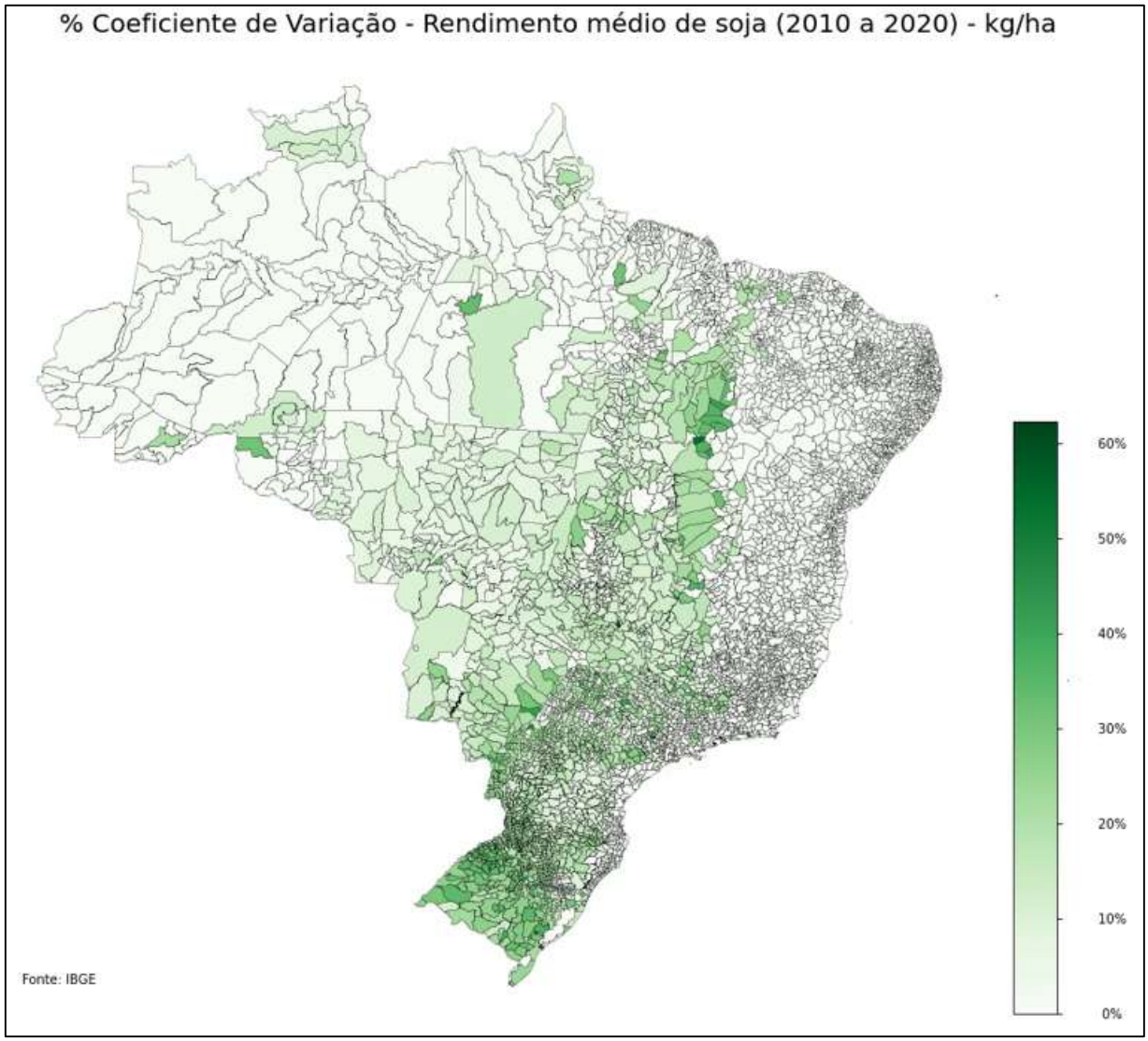

Figura 4: Mapa cloroplético do Brasil com o coeficiente de variação do Rendimento médio de soja (2010 a 2020) - kg/há por município

Observa-se pela figura 1 - Mapa cloroplético do Brasil por Rendimento médio de soja (2010 a 2020) - kg/há, que os estados que apresentam rendimentos médios mais alto são os da região sul e centro-sul do país.

Quando extrapolamos essa análise por município, conforme demonstrado na figura 2, observa-se uma concentração das maiores produtividades dentro de cada estado. Corroborado com a figura 3 escala o dado de acordo com o coeficiente de variação do rendimento médio, ou seja, de acordo com a sua volatilidade. Devido ao tamanho do Brasil e suas particularidades de bioma, o clima ser diferente em cada região, havendo também microclimas dentro de um mesmo estado. Um dos sinônimos desse feito é observado através das diferenças de produtividades de soja, o que tem uma forte dependência do clima para ser favorável.

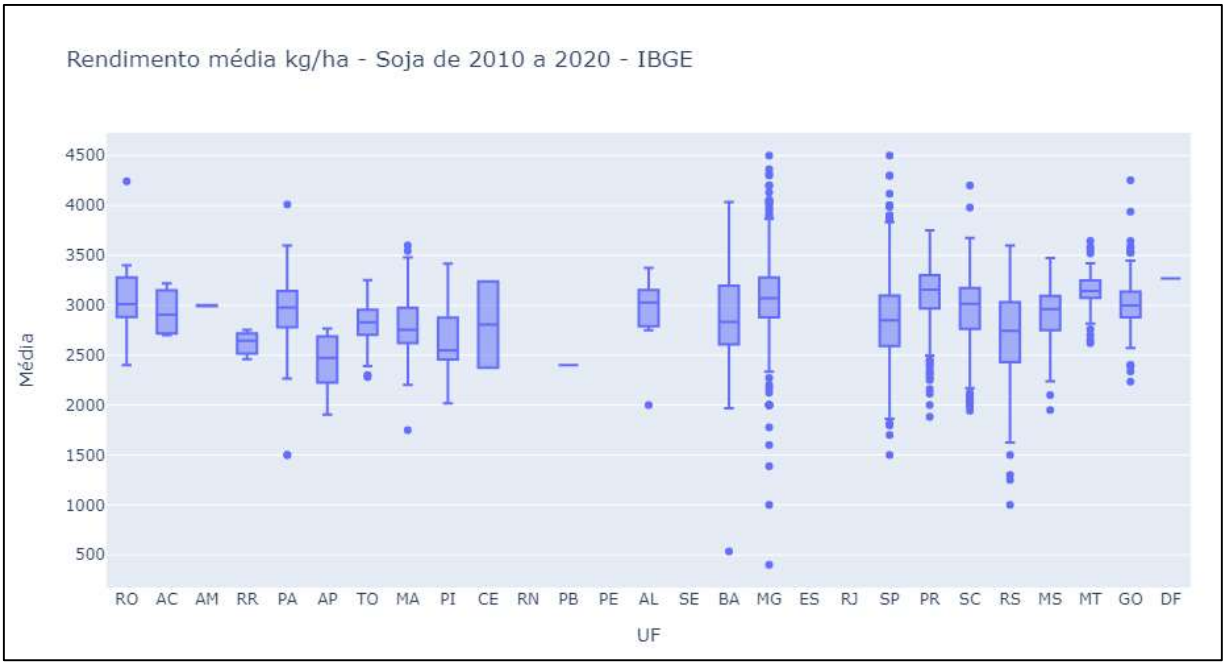

Figura 5: BoxPlot - Rendimento médio kg/ha de soja 2020 - IBGE - por UF

Através BoxPlot é possível observar diferentes comportamentos dos dados de produtividades médias entres as unidades federativas. Para o estado do Rio Grande do Sul observa-se uma elevada variabilidade dos dados de produtividade, o que é corroborado pela grande variabilidade pluviométrica observada entre os anos acarretando episódios frequentes de seca e influenciando diretamente nos rendimentos médios por hectare do estado. No Mato Grosso, em contrapartida, onde o índice pluviométrico é mais regular, a produtividade expressa no BoxPlot demonstra uma maior regularidade dos dados.

Reforçando a intenção deste trabalho em identificar similaridade entre os municípios de cada estado e consequentemente agrupar os melhores indicadores vislumbrando uma dispersão geográfica do risco e/ou adequação das condições; cobertura e taxas, por exemplo; em função do risco embutido.

Área plantada de soja em hectare em 2020:

A área plantada por cultura e regiões geográficas do Brasil (Unidades da Federação, Mesorregiões Geográficas, Microrregiões Geográficas e Municípios) é fornecida pelo Instituto Brasileiro de Geografia e Estatística (IBGE) anualmente. Os dados de soja considerados nesse estudo foram extraídos do site Sistema IBGE de Recuperação Automática (SIDRA) na aba Produção Agrícola Municipal (PAM) (https://sidra.ibge.gov.br/pesquisa/pam/tabelas). Para trabalhar o dataframe importado do SIDRA, é necessário ajustar a formatação e tipo das variáveis.

Para avaliar a representatividade dos municípios em termos de área plantada para a cultura de soja, para a clusterização foi considerado apenas a informação do último ano publicado, 2020, no site do SIDRA.

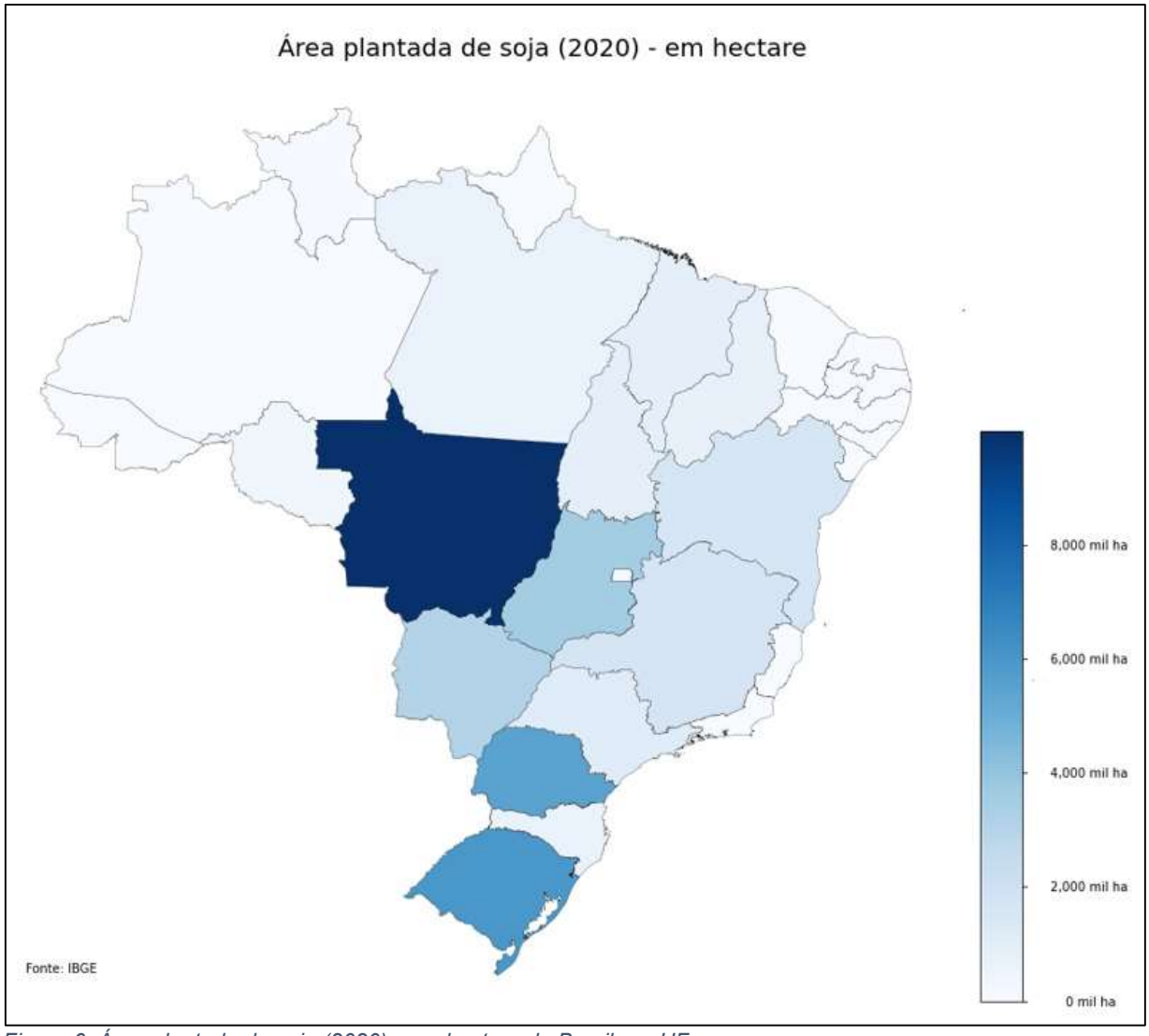

Figura 6: Área plantada de soja (2020) - em hectare do Brasil por UF

Quanto a área plantada de soja de 2020 por estado segue uma distribuição diferente quando comparado com as maiores médias de rendimento médio por estado. O estado com maior representatividade em termos de área plantada segundo o IBGE em 2020 é o Mato Grosso, seguido do Rio Grande do Sul e Paraná.

Sinistralidade da cultura de soja de 2010 a 2020 e Área segurada 2021:

A sinistralidade da cultura de soja de 2010 a 2020 se refere a relação entre as indenizações pagas pelas Seguradoras e os prêmios pagos pelos Segurados nesse período para as apólices contempladas pelo Programa de Subvenção ao Prêmio o Seguro Rural (PSR). Essa informação traz a sensibilidade dos municípios com maior propensão de quebra de safra em função do clima, considerando que a maioria dos produtos de seguro do mercado visam mitigar os efeitos das adversidades climáticas.

Quanto a área segurada de 2021, nos norteia a respeito do apetite de risco, cultura de contratação de seguro etc. Para maiores informações a respeito do programa, recomenda-se a leitura do seguinte artigo: https://www.gov.br/agricultura/pt-br/assuntos/riscos-seguro/seguro-rural/dados/relatorios/relatorio-geral-psr-2021-final.pdf.

Para baixar os dados públicos do site do Ministério da Agricultura, Pecuária e Abastecimento (MAPA) foi necessário utilizar uma biblioteca do Python chamada “urllib3” que permitiu se conectar ao site e para realizar o download das informações foi utilizada a

biblioteca “wget”. Tornando não só o acesso e download das informações possível, como também o tratamento e análise dos dados em função do tamanho dos arquivos.

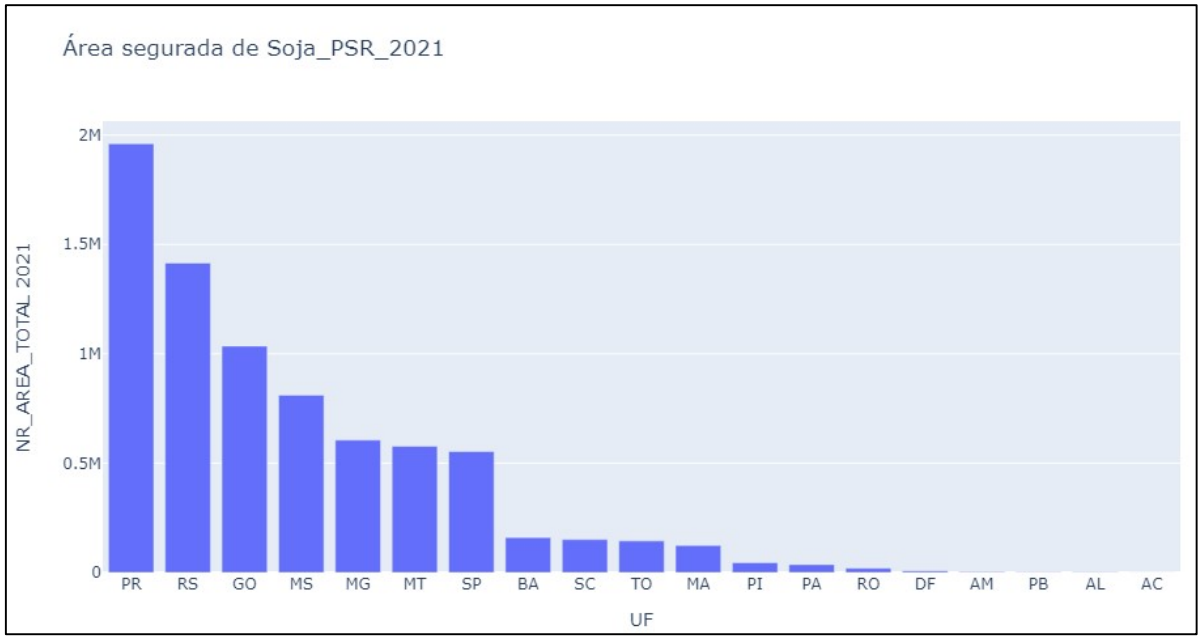

Figura 7: Área Segurada 2021_Soja_PSR

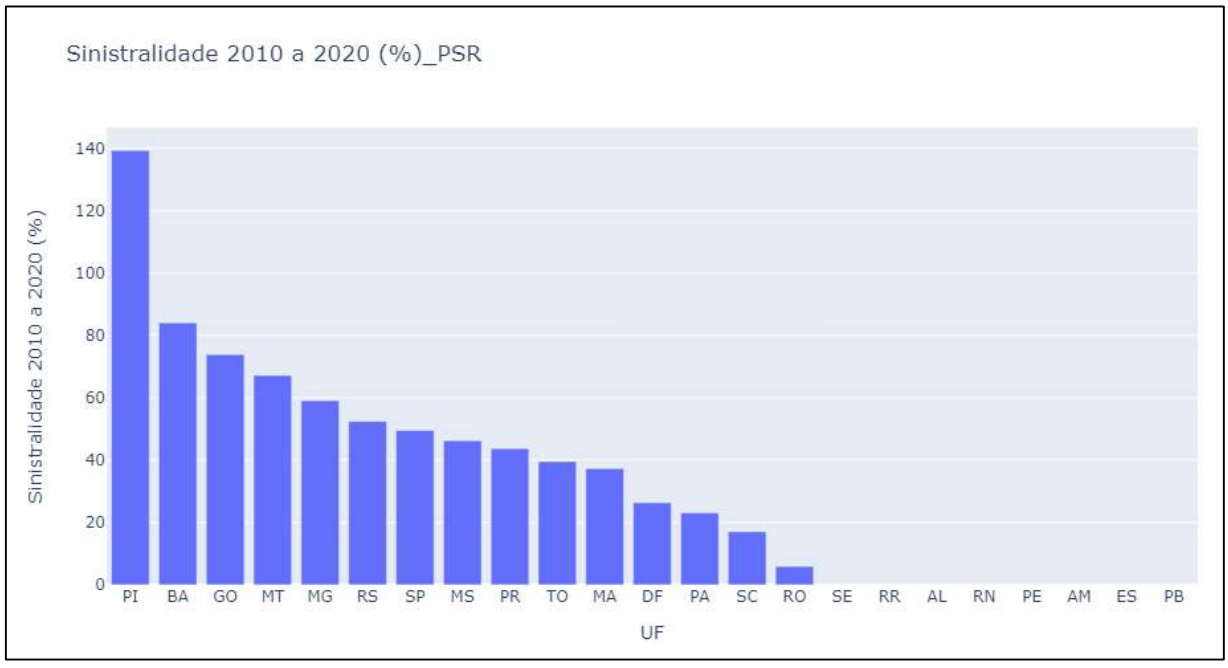

Figura 8: Sinistralidade de 2010 a 2020_Soja_PSR

No gráfico de barras da área segurada em 2021 segundo o PSR, observamos que o estado do Mato Grosso apesar de ser o estado que mais produz soja no país, não é o estado que mais contrata seguro agrícola, possivelmente em função do índice pluviométrico ser mais homogêneo no estado e em função dos produtos que estão disponíveis no mercado não atenderem de forma tão aderente a realidade da região. Vale ressaltar que mesmo que houvesse procura de contratação na mesma proporção da área produtiva não haveria capacidade no mercado para absorver tal demanda, de todo modo, esse não é o tema do trabalho em questão.

Quanto a sinistralidade, observamos que apesar do estado do Paraná, por exemplo, ser o estado que mais demande contratação de Seguro agrícola, a sinistralidade é mais baixa do que outros estados em função da dispersão de risco nele mesmo, sem contar com o avanço da tecnologia por se tratar de uma das regiões mais antigas de produção de grãos.

Dados de reanálise de precipitação por município disponibilizado pelo CHIRPS:

No caso específico deste projeto, foram extraídos dados de precipitação total acumulada mensal de satélite para todos os municípios brasileiros de janeiro de 1981 a junho de 2022 a partir das coordenadas geográficas de seus respectivos centroides e através dos pacotes Geopandas e Raster do Python. A partir destes dados, selecionaram-se os anos completos de 1992 a 2021 para obtenção da Normal Climatológica de cada município. A Normal Climatológica de uma determinada variável climatológica é a média de 30 anos desta variável para um determinado período do ano, em particular os meses do ano. As Normais são utilizadas como referência para avaliação de eventos climáticos, além de fornecerem um contexto anual de variabilidade, para uma determinada região.

Além das Normais Climatológicas, foram obtidas para cada município brasileiro os desvios-padrões e coeficientes de variação dos índices pluviométricos mensais.

Finalmente, como o interesse era a produtividade da safra de grão de verão, foram selecionados, para a clusterização, apenas os meses de setembro a março, período em que se compreende a safra de grãos de verão no Brasil. Na realidade, o período total da safra de soja (plantio, desenvolvimento da planta e colheita) é um pouco mais curto, levando entre 4 a 5 meses para se completar. No entanto, o Brasil é um país de dimensões continentais, com variações climáticas que fazem com que as regiões tenham diferentes inícios, durações e finais de safra. Além disso, dentro de uma mesma região pode haver algumas variações, tanto no início quanto na duração do período de safra.

O mapa de coeficiente de variação médio de precipitação acumulada considerando os últimos 30 anos, corrobora com as variabilidades avaliadas a título de rendimento médio. Ou seja, observando pelo prisma de volatilidade de índice pluviométrico, a produtividade apresenta alta simetria com essa informação, à medida que uma região apresenta maior probabilidade de ocorrência de seca o rendimento médio tem a tendência de acompanhar tão regime.

Como exemplo, podemos citar novamente o estado do Rio Grande do Sul onde notamos no mapa o alto coeficiente de variação de pluviometria corroborando com a alta volatilidade do rendimento médio observado no gráfico do BoxPlot.

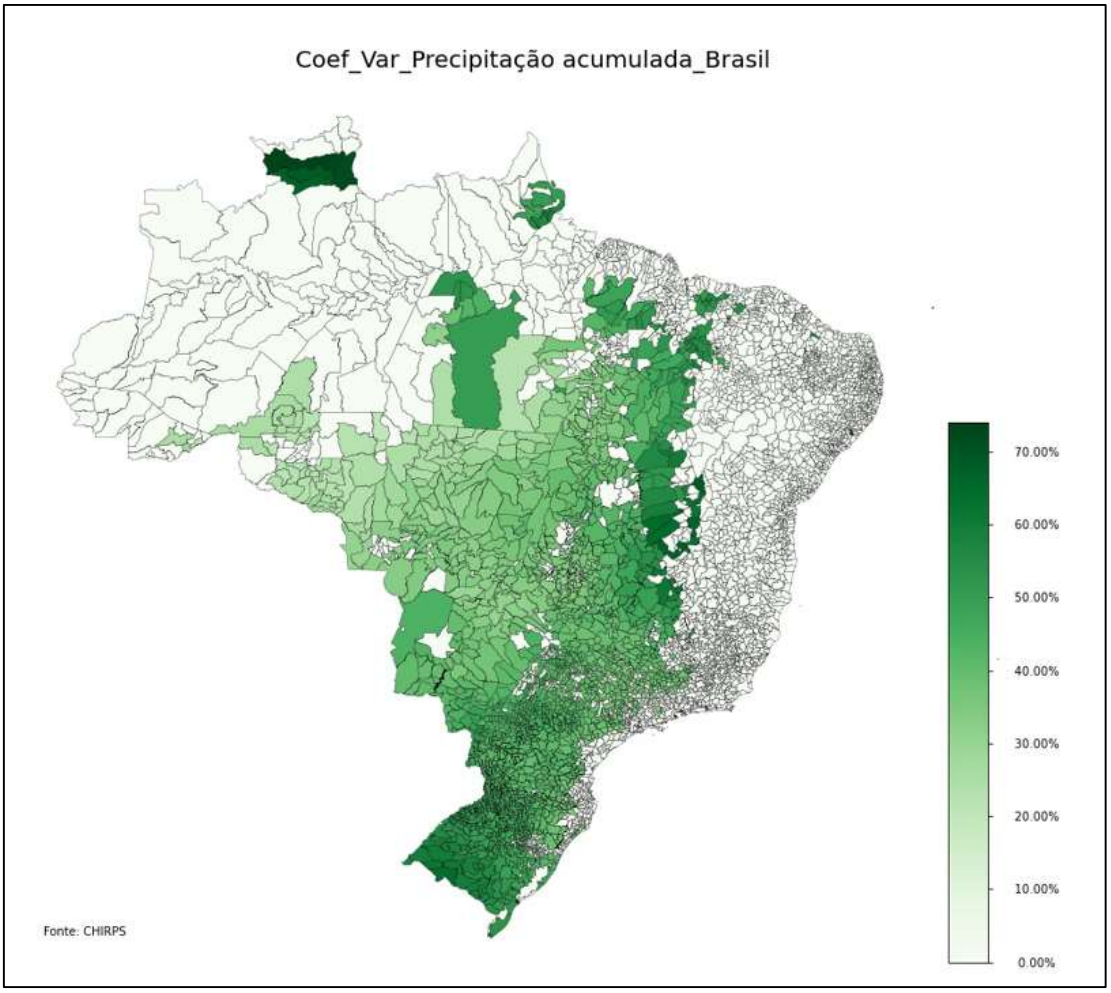

Figura 9: Coef_Var_Precipitação acumulada_Brasil

5 Aplicação da metodologia e resultados

Para aplicação da metodologia descrita acima, foram feitos alguns tratamentos nos dados, elencados a seguir:

- Foram desconsiderados os municípios cujo a área plantada em 2020 segundo do IBGE foi inferior a 100 hectares;
- Foram desconsiderados os municípios com dados de produtividade ausentes em 2018, 2019, 2020;
- Os dados ausentes de produtividade ausentes foram completados com a média dos últimos 10 anos e as médias de produtividade dos municípios foram calculadas com base nos últimos 10 anos (2010 a 2020) após o ajuste citado;
- Foram considerados apenas os estados mais relevantes em termos de cultivo de soja (BA, GO, MA, MG, MS, MT, PI, PR, RS, SC, SP e TO);
- Na clusterização foram consideradas as variáveis listadas abaixo a fim de se identificar similaridade e agrupar municípios com perfis semelhantes:
    - Área plantada em 2020;
    - Média de produtividade de cada município de 2010 a 2020;
    - Coeficiente de variação de produtividade de 2010 a 2020;
    - Sinistralidade de 2010 a 2020;
    - Média de precipitação acumulada dos últimos 30 anos (1992 a 2021) por mês que compõe a safra de verão de um modo geral (setembro, outubro, dezembro, janeiro, fevereiro e março). Ou seja, foram consideradas sete variáveis para cada município;
    - Coeficiente de variação de precipitação acumulada dos últimos 30 anos (1992 a 2021) por mês que compõe a safra de verão de um modo geral (setembro, outubro, dezembro, janeiro, fevereiro e março). Ou seja, foram consideradas sete variáveis para cada município;
- Os clusteres foram feitos por estado e depois avaliado o resultado de cada estado a fim de se identificar um cluster “ótimo”, ou seja, onde as variáveis se comportam de forma semelhante e com indicadores de que tal grupo possui elevado potencial produtivo ou ao menos promissor;
- Para definição dos clusteres de cada estado foi utilizada o método do silhouette.

Trata-se de um índice para a seleção do número de cluster que se dá através da atribuição de um número de cluster máximo a fim de se definir um range que o algoritimo irá percorrer para localizar o que melhor representa a quantidade de clusteres que expressa aquele conjunto de dados. Desta forma, é feito a média de silhoutte de cada número de cluster e escolhido o que número de cluster que possui a maior média. Ou seja, quanto menor for a distância das variáveis analisadas dentro do cluster e quanto maior for a distância em relação aos clusteres vizinhos, melhor será o silhouette. Indicando que a similaridade entre os clusteres.

Esta análise foi realizada para cada estado com a intenção de que cada um tivesse o seu número de cluster mais aderente.

Como as variáveis explicativas possuem escalas diferentes, antes de aplicar a metodologia foi necessário realizar a normalização dos dados.

Como exemplo, segue o gráfico de silhouette para o estado de Minas Gerais no qual o algoritmo identificou como 2 sendo o número de cluster ótimo, que foi o que obteve a maior média dos silhouettes.

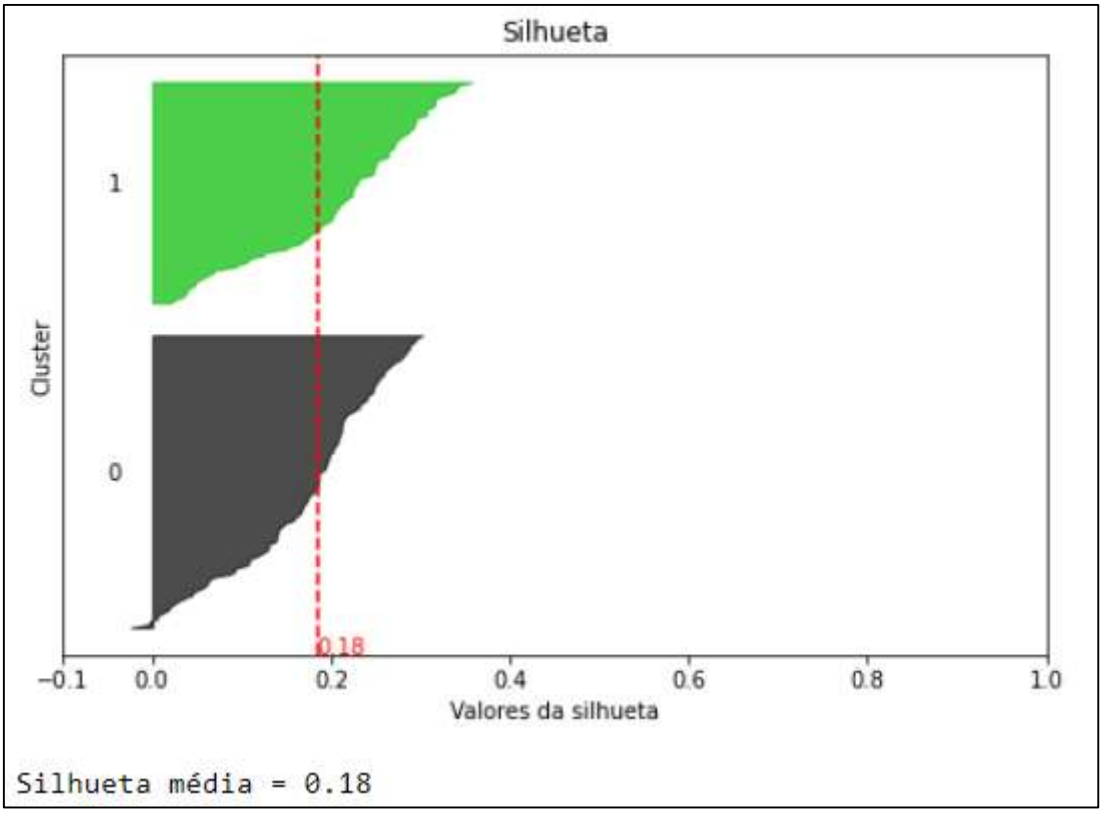

Figura 10: Gráfico de Silhueta de MG

5.1 Análise de resultados

A avaliação do cluster que apresenta um maior potencial produtividade de soja com uma menor volatilidade por estado se deu através da avaliação da maior produtividade média, atrelada ao menor coeficiente de variação médio e considerando a representatividade dos municípios que compõe o cluster tomando como base a áreas plantadas em 2020 segundo o IBGE. Conhecimentos prévios com a cultura de soja e/ou mercado de seguro agrícola também foram utilizadas nessa análise a fim de que mesmo os clusteres sendo calculados automaticamente atrás da aplicação de aprendizado não supervisionado, a interpretação e avaliação levasse em conta experiências prévias vislumbrando inclusive uma possibilidade de calibração da ferramenta. Sabemos que a ciência de dados não tem por finalidade excluir a avaliação de um analista, mas sim possibilitar insights para uma tomada de decisão mais assertiva.

Na figura abaixo, é possível verificar um resumo descritivo dos clusteres por estado, porém, é importante ressaltar que para atingir esse resultado foi necessário realizar tratamento aos dados considerados como informado nas seções anteriores deste trabalho, em função disso, notar-se-á que o somatório de municípios relacionados a cada Unidade da Federação não corresponde ao valor total real. Além disso, é importante frisar que apesar da sinistralidade ter sido incluída na tabela como uma medida ilustrativa, ela não corresponde a todos os municípios considerados em cada cluster, uma vez que pode não ter ocorrido contratação de seguro e consequentemente elegibilidade ao PSR (já que a sinistralidade deriva dos números publicado no Programa de Subvenção ao Seguro Rural).

Ainda sobre a figura abaixo, é possível observar que número de cluster ótimo obtido através do método de silhouette para a maioria dos estados foi igual a dois.

Como era de se esperar, os estados mais representativos em termos de área plantada possuem a maior concentração de municípios, como por exemplo, Mato Grosso, Rio Grande do Sul, Paraná e Goiás.

UF	kmeans	Área plantada_2020	Contagem_Municípios	Produtividade_Média_Ponderada	% CV Prod. Méd. Ponderado	Sinistralidade ponderada 2010 a 2020 (%)
BA	0	129.150	10	2.742,71	21.34	52.14
	1	1.493.150	7	3.079,57	20.02	86.62
GO	0	2.407.200	147	3.137,71	12.33	77.66
	1	1.162.840	47	3.107,50	12.32	63.39
MA	0	791.269	35	2.780,40	17.88	38.30
	1	165.238	22	2.996,22	12.36	37.76
MG	0	1.151.512	174	3.104,13	13.86	55.95
	1	541.028	39	3.117,12	14.84	68.87
MS	0	763.977	23	3.204,58	11.76	53.80
	1	2.356.895	52	3.045,77	15.66	44.98
MT	0	4.709.922	71	3.146,55	7.21	80.87
	1	5.275.552	47	3.190,83	8.27	49.73
PI	0	15.615	7	2.840,28	17.97	0.00
	1	727.490	18	2.626,19	27.17	134.42
PR	0	3.522.240	259	3.201,55	14.66	46.30
	1	2.010.597	119	3.312,47	16.37	37.53
RS	0	2.049.884	123	2.256,34	26.20	71.71
	1	3.945.023	280	2.875,23	25.29	39.09
SC	0	362.516	84	3.270,02	12.56	18.68
	1	300.157	89	3.110,19	18.35	15.09
SP	0	838.094	177	3.118,73	17.73	46.44
	1	280.236	187	2.902,40	16.07	53.95
TO	0	398.230	37	2.848,00	10.14	38.26
	1	429.556	30	2.880,96	12.65	38.62
	2	130.023	14	2.916,56	7.27	71.26

Figura 11: Tabela resumo K-Means por UF

A seguir as análises foram divididas por estado para que fosse possível identificar padrões em subgrupos e definir um potencial cluster “ótimo”.

Bahia:

O cluster que apresentou um melhor resultado analisando a combinação de Produtividade média, coeficiente de variação médio e representatividade em termos de área plantada foi o cluster 1, com maior concentração dos municípios mais produtivos na mesorregião do Extremo Oeste Baiano.

Observa-se que grande parte do estado não foi considerado para clusterização em função dos requisitos estabelecidos para tal e seguindo em linha com a concentração agrícola de produção de soja do estado que é mais fortalecida no Oeste Baiano.

UF	kmeans	Nome_Mesorregião	Área plantada_2020	Contagem_Municípios	Produtividade_Média_Ponderada	% CV Prod. Méd. Ponderado	Sinistralidade ponderada 2010 a 2020 (%)
BA	0.0	Centro Sul Baiano	600	10	2.619,00	15.98	NaN
		Extremo Oeste Baiano	123.900	10	2.715,58	21.77	52.14
		Vale São-Franciscano da Bahia	4.650	10	3.481,73	10.52	NaN
	1.0	Extremo Oeste Baiano	1.493.150	7	3.079,57	20.02	86.62

Figura 12: Tabela resumo K-Means por Mesorregião_BA

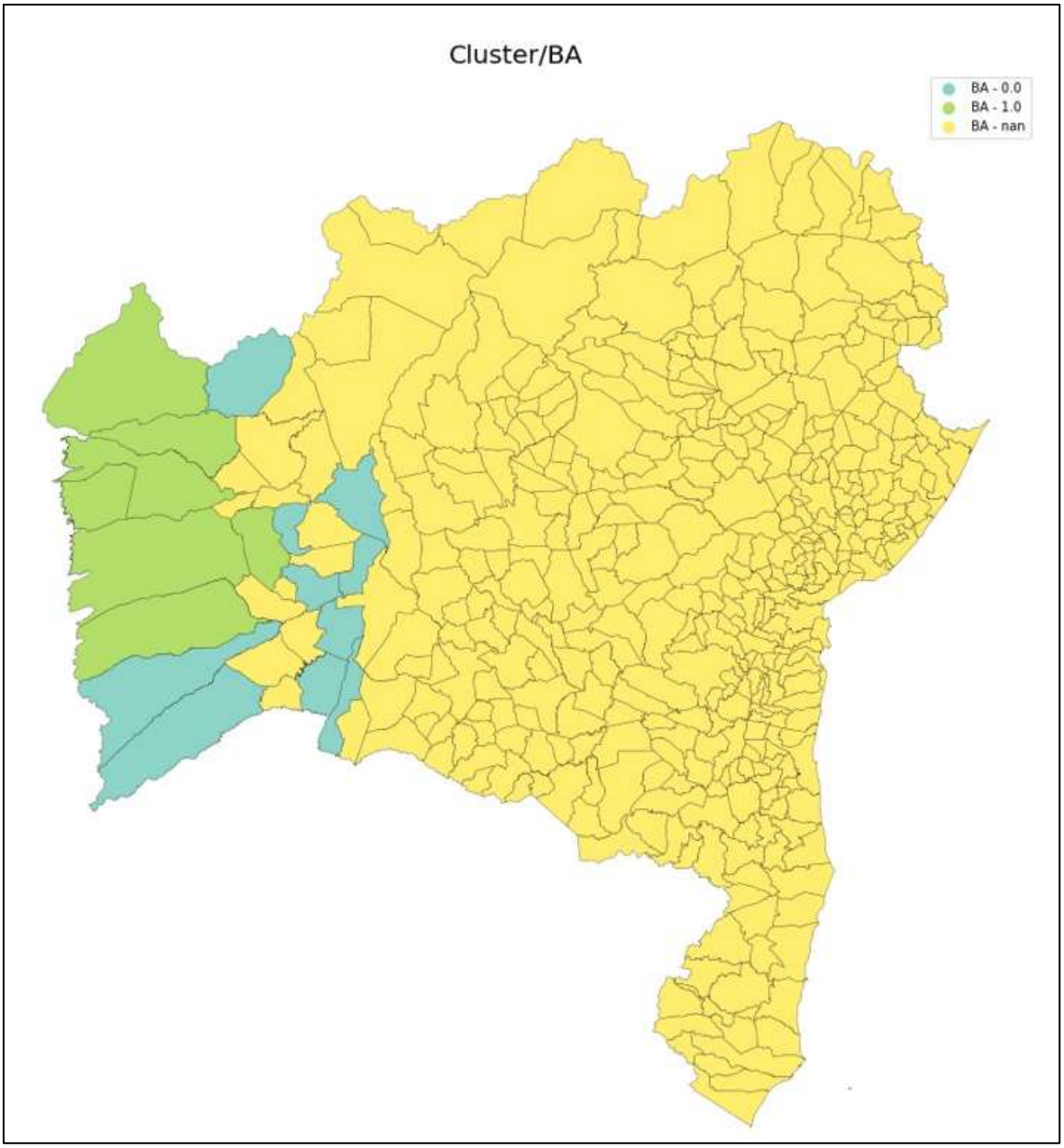

Figura 13: Mapa K-Means_BA

	UF	Geocódigo	Município (UF)	kmeans	Área plantada_2020	Média Prod 2010 a 2020
0	BA	2911105	Formosa do Rio Preto (BA)	1	427.500,00	3.114,45
1	BA	2928901	São Desidério (BA)	1	384.400,00	3.110,64
2	BA	2903201	Barreiras (BA)	1	195.500,00	3.185,64
3	BA	2909307	Correntina (BA)	1	193.100,00	2.833,45
4	BA	2919553	Luís Eduardo Magalhães (BA)	1	162.200,00	3.225,18
5	BA	2926202	Riachão das Neves (BA)	1	116.500,00	2.945,91
6	BA	2902500	Baianópolis (BA)	1	13.950,00	2.498,00

Figura 14: Rank dos principais municípios do cluster escolhido_BA

Goiás:

Para este estado o coeficiente médio de variação não apresentou resultado significativo entre os clusteres, possivelmente, seria necessário subdividir de forma manual em mais

grupos para então avaliarmos se o comportamento de tal variável permanece. Em termos de produtividade média, o cluster 0, obteve resultado ligeiramente melhor, e curiosamente é o cluster com área segurada total maior. Levando em consideração o exposto e sem realizar ajuste manual na definição dos clusteres, o cluster 0 apresentou um melhor resultando considerando a combinação das variáveis analisadas.

O cluster 0 concentrou grande parte das áreas cultivadas de soja da mesorregião Sul Goiano, onde de fato a produção agrícola é mais desenvolvida neste estado.

UF	kmeans	Nome_Mesorregião	Área plantada_2020	Contagem_Municípios	Produtividade_Média_Ponderada	% CV Prod. Méd. Ponderado	Sinistralidade ponderada 2010 a 2020 (%)
GO	0.0	Centro Goiano	137.862	147	3.007,78	8.93	52.71
		Leste Goiano	49.324	147	2.955,87	8.57	46.24
		Noroeste Goiano	153.159	147	3.089,89	13.49	66.44
		Norte Goiano	88.645	147	2.965,60	16.13	49.63
		Sul Goiano	1.978.210	147	3.162,71	12.40	82.09
	1.0	Leste Goiano	524.607	47	3.083,69	12.13	53.72
		Norte Goiano	104.013	47	3.039,27	12.66	69.05
		Sul Goiano	534.220	47	3.144,17	12.44	68.82

Figura 15: Tabela resumo K-Means por Mesorregião_GO

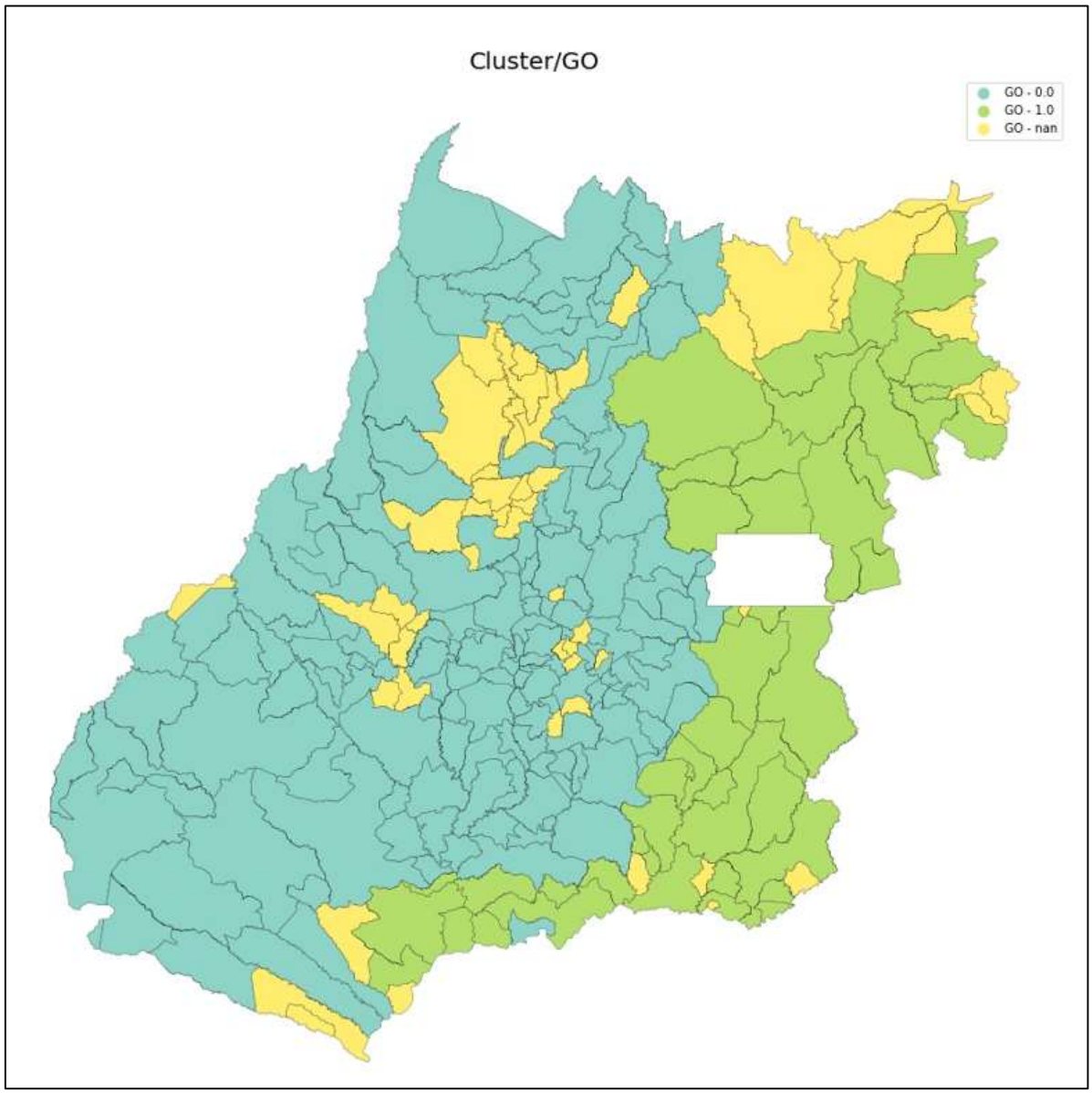

Figura 16: Mapa K-Means_GO

	UF	Geocódigo	Município (UF)	kmeans	Área plantada_2020	Média Prod 2010 a 2020
0	GO	5218805	Rio Verde (GO)	0	390.000,00	3.110,91
1	GO	5211909	Jataí (GO)	0	285.000,00	3.346,55
2	GO	5213756	Montividiu (GO)	0	138.000,00	3.189,09
3	GO	5216403	Paraúna (GO)	0	115.000,00	3.216,36
4	GO	5213103	Mineiros (GO)	0	100.000,00	3.244,45
5	GO	5205471	Chapadão do Céu (GO)	0	90.000,00	3.261,00
6	GO	5209101	Goiatuba (GO)	0	75.800,00	2.937,73
7	GO	5220603	Silvânia (GO)	0	72.000,00	3.343,64
8	GO	5217104	Piracanjuba (GO)	0	68.000,00	3.175,27
9	GO	5219308	Santa Helena de Goiás (GO)	0	51.000,00	3.181,82

Figura 17: Rank dos principais municípios do cluster escolhido_GO

Maranhão:

O cluster 1 atenderia as premissas de maior produtividade média de soja e menor coeficiente de variação. Porém, com relação a desenvolvimento agrícola na região o cluster que apresentou maior área plantada neste estado foi o cluster 0. Avaliando a área segurada segundo o PSR em 2021 para os municípios que compõe cada cluster, o cluster 0 teve mais contratação em relação ao cluster 1, em torno de 78,3 mil hectares versus 36,1 mil hectares. Uma outra premissa que chama atenção é que a sinistralidade de 2010 a 2020 segundo o PSR, mesmo não se tratando de uma variável que abarque todos os municípios de cada cluster, o resultado é similar nos dois.

Avaliando o coeficiente de variação médio de precipitação acumulada para os meses que contemplam a safra de soja no geral (setembro a março) dos últimos 30 anos, o cluster 0 apresenta um menor resultado, o que constrói uma justificativa razoável em considerar o cluster 0 como o cluster "ótimo" desse estado dentre os calculados pelo algoritmo atrelado à disposição geográfica do mesmo, que faz parte da fronteira agrícola do MATOPIBA (Maranhão, Tocantins, Piauí e Bahia) e que está em expansão e desenvolvimento, e considerando neste caso que tal cluster possui uma maior relevância em termos de concentração de área plantada. Ficando a recomendação no caso de um maior apetite de risco por essa região em função da classificação como "cluster ótimo", a oferta de produtos e precificação adequada considerando as particularidades em termos de volatilidade.

		CV_Med_Precip_(%)
UF	kmeans	
MA	0	47.20
	1	52.73

Figura 18: Tabela resumo de CVAR Médio de precipitação_K-Means MA

UF	kmeans	Nome_Mesorregião	Área plantada_2020	Contagem_Municipios	Produtividade_Média_Ponderada	% CV Prod. Méd. Ponderado	Sinistralidade ponderada 2010 a 2020 (%)
MA	0.0	Centro Maranhense	24.088	35	2.892,66	13.53	30.68
		Leste Maranhense	40.829	35	2.899,26	16.77	27.81
		Oeste Maranhense	63.834	35	3.131,93	12.90	0.00
		Sul Maranhense	662.518	35	2.735,12	18.59	41.24
	1.0	Centro Maranhense	475	22	2.937,67	16.05	NaN
		Leste Maranhense	90.873	22	2.659,32	17.66	41.04
		Oeste Maranhense	73.890	22	3.410,93	5.81	0.00

Figura 19: Tabela resumo K-Means por Mesorregião_MA

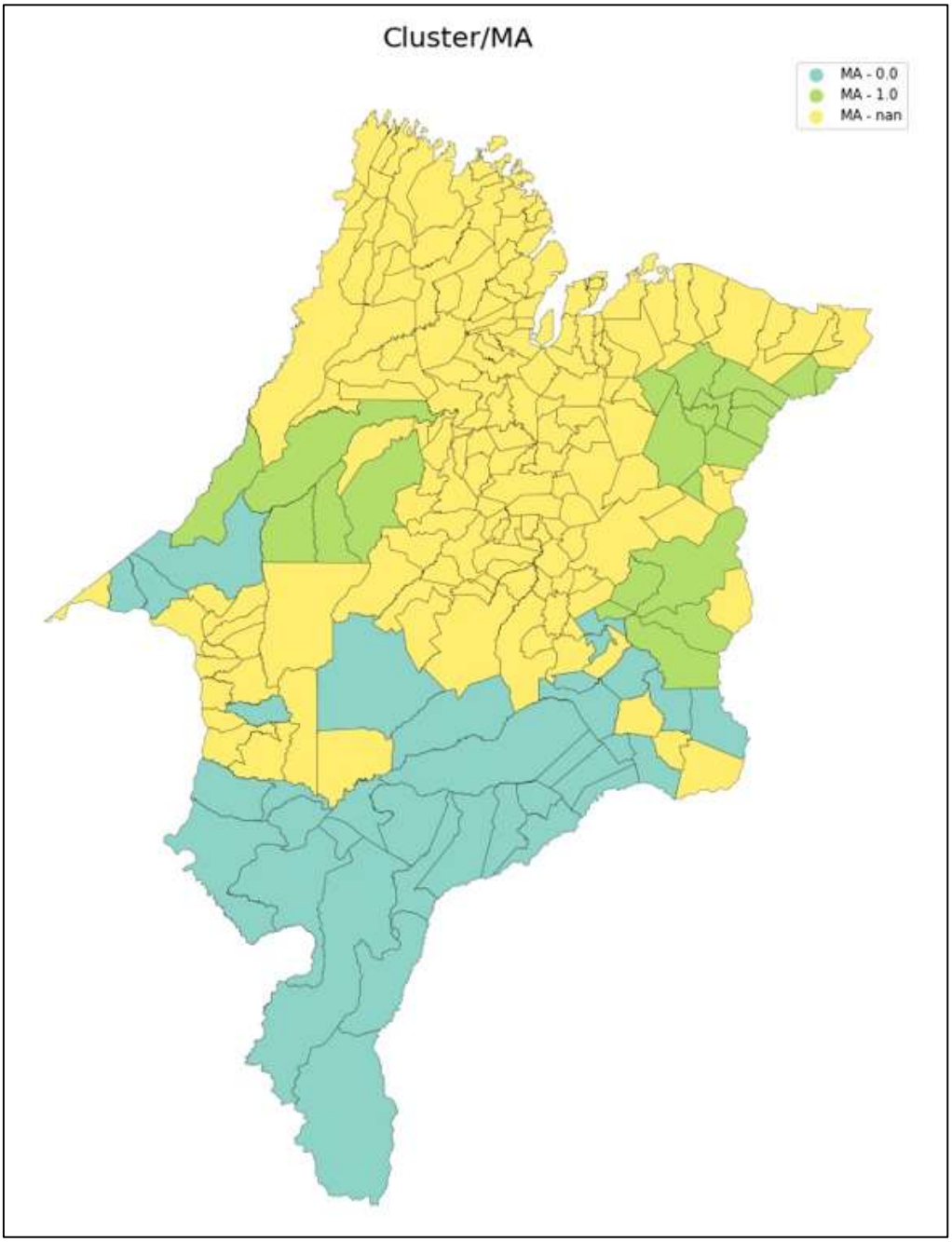

Figura 20: Mapa K-Means_MA

	UF	Geocódigo	Município (UF)	kmeans	Área plantada_2020	Média Prod 2010 a 2020
0	MA	2101400	Balsas (MA)	0	195.161,00	2.754,36
1	MA	2112001	Tasso Fragoso (MA)	0	188.250,00	2.752,82
2	MA	2100501	Alto Parnaíba (MA)	0	54.035,00	2.771,73
3	MA	2100055	Açailândia (MA)	0	53.179,00	3.159,80
4	MA	2109502	Riachão (MA)	0	52.500,00	2.690,18
5	MA	2109700	Sambaíba (MA)	0	44.500,00	2.699,45
6	MA	2106102	Loreto (MA)	0	35.550,00	2.622,91
7	MA	2110658	São Domingos do Azeitão (MA)	0	28.387,00	2.812,00
8	MA	2102804	Carolina (MA)	0	26.500,00	2.653,27
9	MA	2111607	São Raimundo das Mangabeiras (MA)	0	17.550,00	2.662,55

Figura 21: Rank dos principais municípios do cluster escolhido_MA

Minhas Gerais:

No caso de Minas Gerais a produtividade média ficou similar nos dois clusteres. Quanto ao coeficiente de variação, que traduz de certa forma a volatidade da produtividade dos últimos dez anos, o cluster 0 apresentou um resultado ligeiramente melhor, o que somado a representatividade de tal cluster no quesito área plantada, o torna um cluster com municípios com potencial produtividade promissores e similares.

Em termos de mesorregião, o cluster 0, concentra de forma absoluta o Triângulo Mineiro/Alto Paranaíba e o Sul/Sudeste de Minas, onde o cultivo de soja é mais latente.

UF	kmeans	Nome_Mesorregião	Área plantada_2020	Contagem_Municípios	Produtividade_Média_Ponderada	% CV Prod. Méd. Ponderado	Sinistralidade ponderada 2010 a 2020 (%)
MG	0.0	Campo das Vertentes	34.200	174	3.054,62	18.99	3.62
		Central Mineira	8.086	174	2.793,95	14.44	3.91
		Metropolitana de Belo Horizonte	450	174	2.792,83	9.05	NaN
		Noroeste de Minas	17.000	174	3.128,18	10.54	25.63
		Oeste de Minas	55.600	174	3.269,73	15.60	14.00
		Sul/Sudoeste de Minas	132.862	174	3.123,08	18.46	50.78
		Triângulo Mineiro/Alto Paranaíba	903.314	174	3.095,51	12.94	58.23
	1.0	Central Mineira	8.300	39	3.077,06	16.11	41.34
		Metropolitana de Belo Horizonte	1.050	39	3.068,71	9.83	0.00
		Noroeste de Minas	475.300	39	3.171,90	14.32	61.66
		Norte de Minas	56.378	39	2.662,01	19.17	249.26

Figura 22: Tabela resumo K-Means por Mesorregião_MG

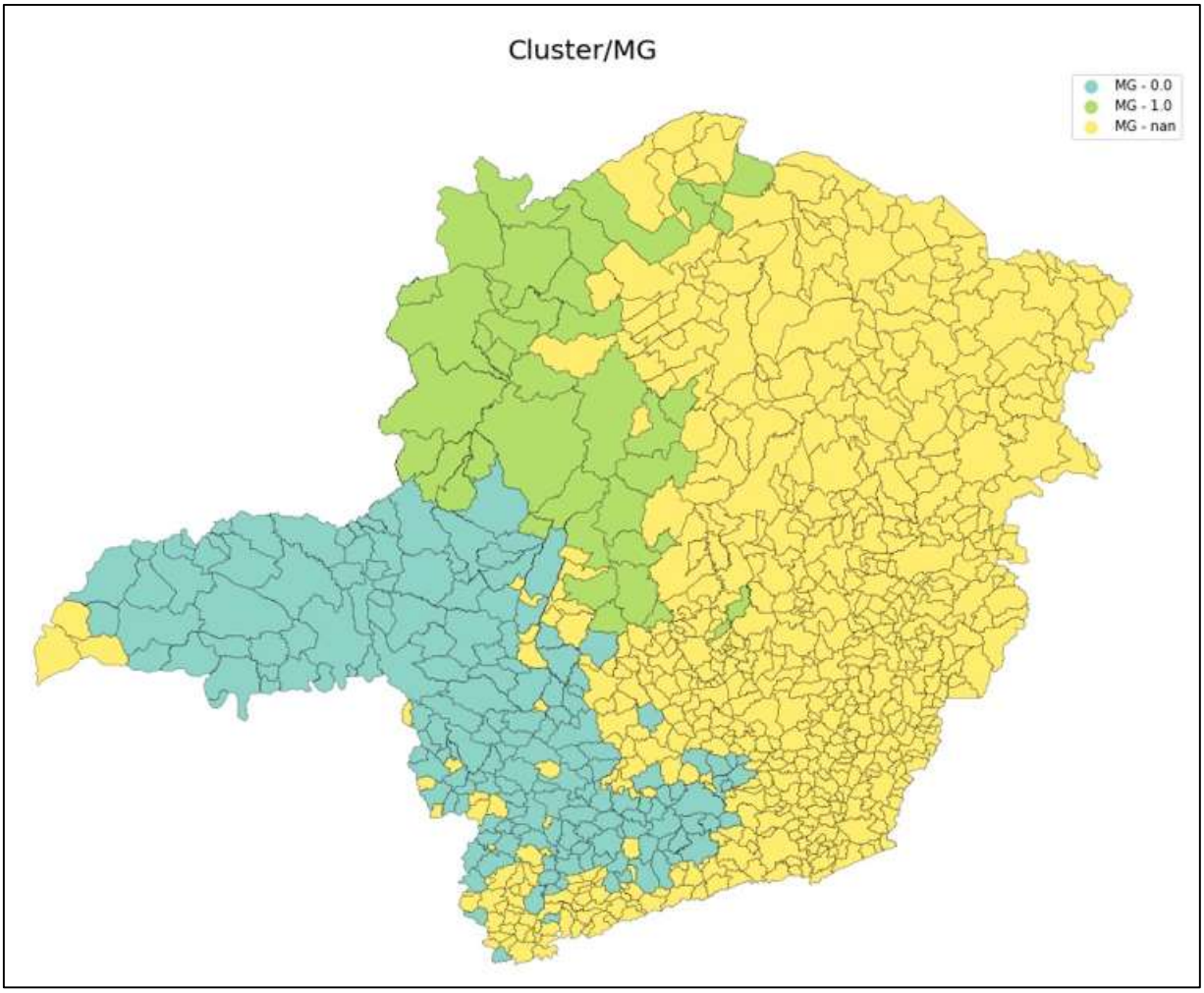

Figura 23: Mapa K-Means_MG

	UF	Geocódigo	Município (UF)	kmeans	Área plantada_2020	Média Prod 2010 a 2020
0	MG	3170107	Uberaba (MG)	0	90.000,00	3.172,73
1	MG	3170206	Uberlândia (MG)	0	60.000,00	3.118,73
2	MG	3119302	Coromandel (MG)	0	58.000,00	3.288,45
3	MG	3129509	Ibiá (MG)	0	40.000,00	3.155,45
4	MG	3156908	Sacramento (MG)	0	40.000,00	2.990,91
5	MG	3149804	Perdizes (MG)	0	38.000,00	3.203,18
6	MG	3117306	Conceição das Alagoas (MG)	0	35.000,00	3.012,73
7	MG	3142809	Monte Alegre de Minas (MG)	0	35.000,00	2.844,55
8	MG	3112604	Capinópolis (MG)	0	32.427,00	3.118,00
9	MG	3148103	Patrocínio (MG)	0	30.000,00	3.148,18

Figura 24: Rank dos principais municípios do cluster escolhido_MG

Mato Grosso do Sul:

Para este estado o cluster 1 apresenta a maior concentração de área, porém a menor produtividade e menor coeficiente de variação em termos de produtividade. Avaliando essas três premissas, as duas que traduzem a produtividade da região corroboram contra a escolha deste cluster como “ótimo”. Além da concentração elevada no sul do estado, no qual apresenta

alguns municípios que demandam atenção maior do ponto de vista agronômico e de manejo devido as características do solo da região.

Em contrapartida a sinistralidade média segundo o PSR de 2010 a 2020 para o cluster 1 foi de 45% enquanto o do cluster 0 foi de 54%.

Sendo assim, o cluster 0 foi considerado como ótimo nesse estudo em função da maioria das variáveis analisadas apresentarem o comportamento esperado para um cluster "ótimo".

UF	kmeans	Nome_Mesorregião	Área plantada_2020	Contagem_Municipios	Produtividade_Média_Ponderada	% CV Prod. Méd. Ponderado	Sinistralidade ponderada 2010 a 2020 (%)
MS	0.0	Centro Norte de Mato Grosso do Sul	476.675	23	3.159,67	13.64	48.04
		Leste de Mato Grosso do Sul	277.829	23	3.288,92	8.63	66.81
		Pantanais Sul Mato-grossense	9.473	23	2.990,91	8.52	102.85
	1.0	Centro Norte de Mato Grosso do Sul	230.000	52	3.075,36	19.69	21.31
		Leste de Mato Grosso do Sul	93.052	52	2.879,97	23.87	72.74
		Pantanais Sul Mato-grossense	31.884	52	2.796,07	18.82	56.64
		Sudoeste de Mato Grosso do Sul	2.001.959	52	3.054,05	14.76	44.63

Figura 25: Tabela resumo K-Means por Mesorregião_MS

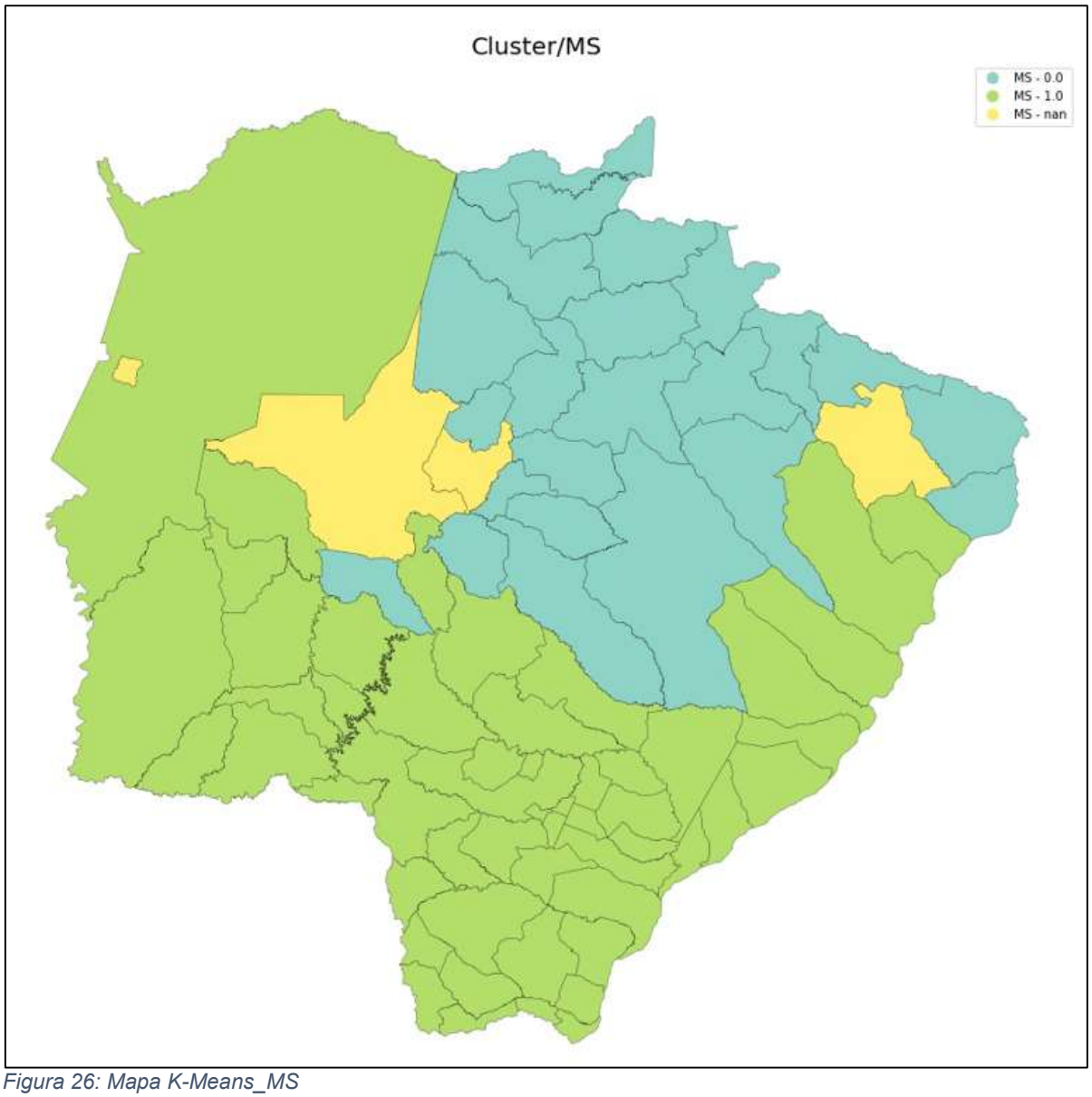

Figura 26: Mapa K-Means_MS

	UF	Geocódigo	Município (UF)	kmeans	Área plantada_2020	Média Prod 2010 a 2020
0	MS	5007695	São Gabriel do Oeste (MS)	0	120.000,00	3.462,73
1	MS	5002951	Chapadão do Sul (MS)	0	91.000,00	3.329,64
2	MS	5002704	Campo Grande (MS)	0	82.000,00	2.956,09
3	MS	5003256	Costa Rica (MS)	0	81.000,00	3.471,73
4	MS	5001508	Bandeirantes (MS)	0	76.000,00	2.970,00
5	MS	5006275	Paraíso das Águas (MS)	0	66.000,00	3.255,00
6	MS	5007935	Sonora (MS)	0	57.300,00	3.217,09
7	MS	5002605	Camapuã (MS)	0	31.030,00	3.076,82
8	MS	5008008	Terenos (MS)	0	28.240,00	3.056,64
9	MS	5007109	Ribas do Rio Pardo (MS)	0	23.886,00	2.754,82

Figura 27: Rank dos principais municípios do cluster escolhido_MS

Mato Grosso:

O cluster 1 apresenta resultado dentro do esperado para concentração relevante de área plantada e produtividade média, com coeficiente de variação um pouco acima do cluster 0. Quando avaliada a sinistralidade de 2010 a 2020 segundo o PSR dos clusteres, o 1 apresenta resultado bem melhor que o cluster 0, este último possui sinistralidade de 81% versus 50% do cluster 1. Quando avaliamos os municípios que o compõe, observamos os municípios com maior produtividade do Brasil como Sorriso/MT, Nova Ubiratã/MT e Nova Mutum/MT. Abaixo segue a lista dos dez maiores municípios em termos de área plantada e produtividade média do cluster escolhido.

UF	kmeans	Nome_Mesorregião	Área plantada_2020	Contagem_Municípios	Produtividade_Média_Ponderada	% CV Prod. Méd. Ponderado	Sinistralidade ponderada 2010 a 2020 (%)
MT	0.0	Centro-Sul Mato-grossense	172.533	71	3.212,13	8.17	102.15
		Nordeste Mato-grossense	1.685.326	71	3.150,47	8.35	95.91
		Norte Mato-grossense	1.282.559	71	3.101,41	7.81	85.00
		Sudeste Mato-grossense	1.358.701	71	3.162,19	5.13	54.39
		Sudoeste Mato-grossense	210.803	71	3.235,48	6.98	83.10
	1.0	Nordeste Mato-grossense	226.250	47	3.254,38	10.45	138.20
		Norte Mato-grossense	5.049.302	47	3.187,98	8.18	43.40

Figura 28: Tabela resumo K-Means por Mesorregião_MT

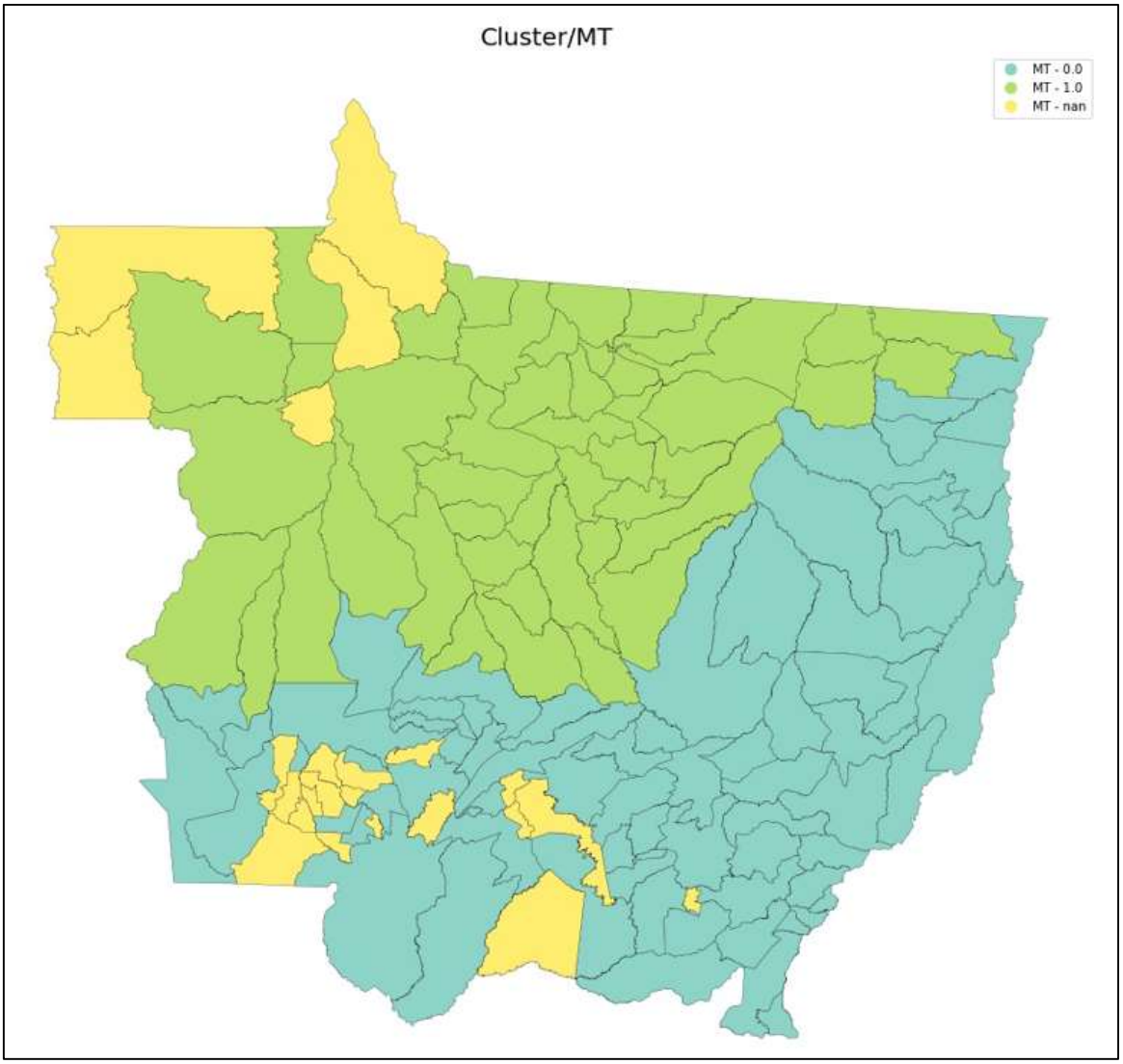

Figura 29: Mapa K-Means_MT

	UF	Geocódigo	Município (UF)	kmeans	Área plantada_2020	Média Prod 2010 a 2020
0	MT	5107925	Sorriso (MT)	1	590.000,00	3.328,27
2	MT	5106240	Nova Ubiratã (MT)	1	396.000,00	3.349,82
3	MT	5106224	Nova Mutum (MT)	1	395.000,00	3.186,00
8	MT	5107875	Sapezal (MT)	1	352.000,00	3.173,09
17	MT	5105259	Lucas do Rio Verde (MT)	1	235.000,00	3.214,00
18	MT	5101902	Brasnorte (MT)	1	230.000,00	3.115,09
20	MT	5104526	Ipiranga do Norte (MT)	1	225.000,00	3.116,00
24	MT	5108907	Nova Maringá (MT)	1	205.000,00	3.090,91
25	MT	5102686	Campos de Júlio (MT)	1	200.500,00	3.101,00
32	MT	5106802	Porto dos Gaúchos (MT)	1	182.000,00	3.172,64

Figura 30: Rank dos principais municípios do cluster escolhido_MT

Piauí:

No Piauí o cluster 1 é o que possui a maior área plantada e maior concentração de seguro, 43 mil hectares versus 2,1 mil hectares do cluster 0. Além disso, também apresenta características de maior volatilidade em termos de produtividade média e menor produtividade comparado com o cluster 0.

No entanto, outras considerações devem ser levadas em conta, como por exemplo:

- A quantidade de municípios que compõe o cluster 0 e que teoricamente poderia levar a interpretação que esse cluster agrupa a maior parte dos municípios com potencial produtivo do estado;

- Avaliando o coeficiente de variação médio de precipitação acumulada para os meses que contemplam a safra de soja no geral (setembro a março) dos últimos 30 anos, o cluster 0 apresenta um maior resultado;

- Outro fator que se deve avaliar é o avanço agrícola, no caso, a mesorregião do Sudoeste Piauiense faz parte da fronteira agrícola do MATOPIBA (Maranhão, Tocantins, Piauí e Bahia) que está em expansão e desenvolvimento.

Os pontos elencados acima nos levam a crer que na necessidade de selecionar um cluster para ser considerado como “ótimo” dentre as opções, o cluster 1 cumpre esse papel, apesar das ressalvas.

		CV_Med_Precip_(%)
UF	kmeans	
PI	0	53.64
	1	50.21

Figura 31: Tabela resumo de CVAR Médio de precipitação_K-Means PI

UF	kmeans	Nome_Mesorregião	Área plantada_2020	Contagem_Municípios	Produtividade_Média_Ponderada	% CV Prod. Méd. Ponderado	Sinistralidade ponderada 2010 a 2020 (%)
PI	0.0	Centro-Norte Piauiense	11.897	7	2.979,58	17.61	0.00
		Norte Piauiense	2.518	7	2.177,47	24.67	NaN
		Sudeste Piauiense	1.200	7	2.850,00	7.44	NaN
	1.0	Sudoeste Piauiense	727.490	18	2.626,19	27.17	134.42

Figura 32: Tabela resumo K-Means por Mesorregião_PI

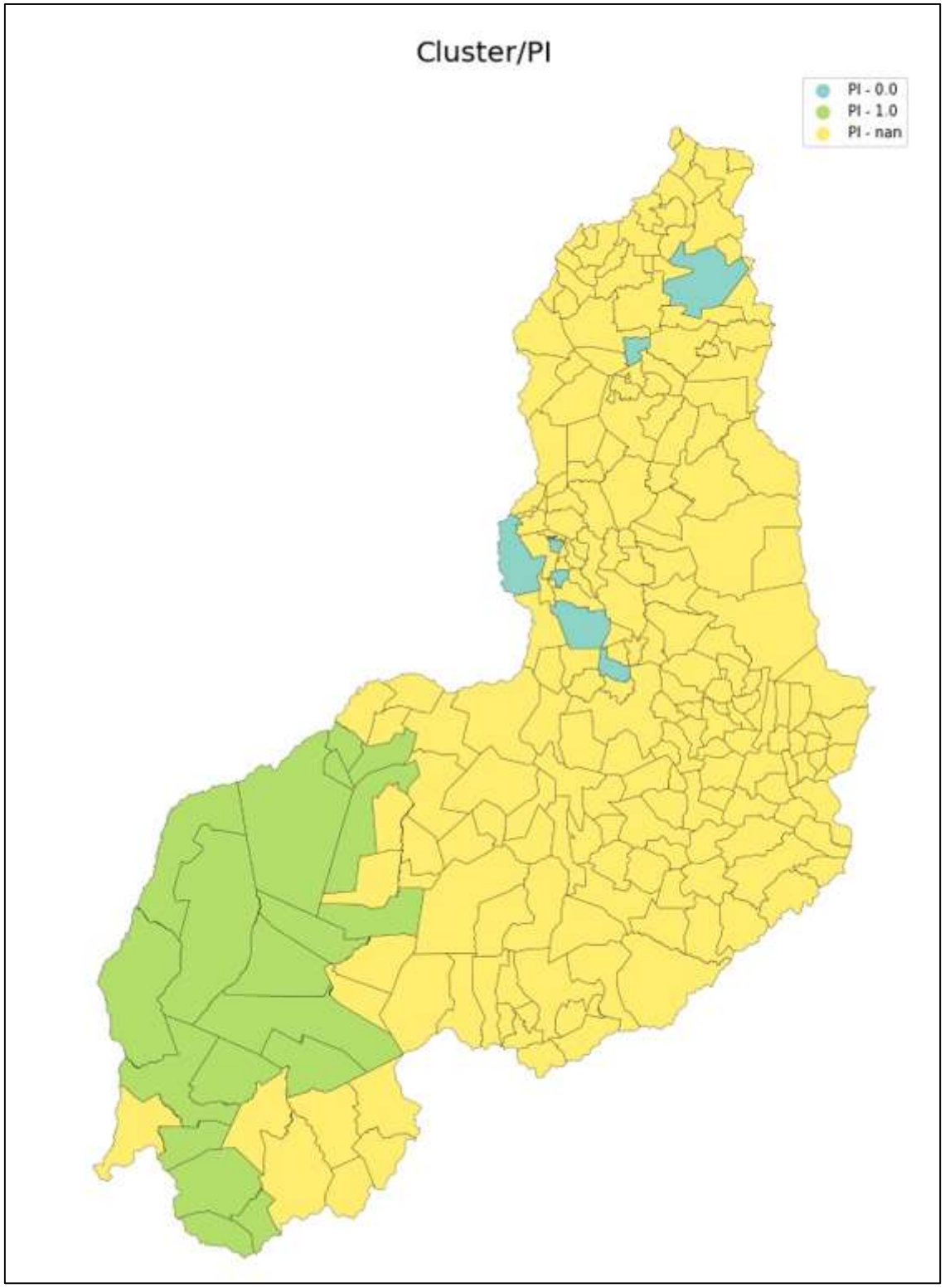

Figura 33: Mapa K-Means_PI

	UF	Geocódigo	Município (UF)	kmeans	Área plantada_2020	Média Prod 2010 a 2020
0	PI	2201150	Baixa Grande do Ribeiro (PI)	1	198.507,00	2.742,45
1	PI	2211209	Uruçuí (PI)	1	146.354,00	2.514,73
2	PI	2208908	Ribeiro Gonçalves (PI)	1	76.622,00	2.969,82
3	PI	2209203	Santa Filomena (PI)	1	71.086,00	2.515,45
4	PI	2201903	Bom Jesus (PI)	1	65.809,00	2.507,36
5	PI	2203230	Currais (PI)	1	45.428,00	2.290,18
6	PI	2204402	Gilbués (PI)	1	29.915,00	2.750,36
7	PI	2210631	Sebastião Leal (PI)	1	27.566,00	2.550,27
8	PI	2206605	Monte Alegre do Piauí (PI)	1	21.877,00	2.616,09
9	PI	2202901	Corrente (PI)	1	20.853,00	2.660,91

Figura 34: Rank dos principais municípios do cluster escolhido_PI

Paraná:

O Paraná é um dos estados que mais possuem a cultura de contratação de seguro agrícola e percepção de utilizá-lo como mais um mecanismo de gerenciamento de risco climático. Desta maneira, um dos desafios é a dispersão de risco dentro do próprio estado, considerando também que algumas regiões ainda possuem microclimas e tipos de solo diferentes que a depender do evento climático impactará de maneiras distintas.

O cluster 0 que abarcou os municípios com maior representatividade em termos de área plantada, também possui o menor coeficiente de variação em relação a produtividade média dos últimos 10 anos e maior concentração de contratação do PSR no ano de 2021 considerando a área plantada, 1,3 milhões de hectares frente a 612 mil hectares do cluster 1. No entanto, a produtividade média do cluster 1 é ligeiramente menor que a do cluster 0.

Os pontos acima já seriam satisfatórios para justificar a escolha do cluster 0 como o cluster “ótimo” dentre os calculados automaticamente pela ferramenta em estudo. Na prática e vislumbrando uma evolução do tema, alguns ajustem poderiam ser testados a fim de calibrar o resultado do algoritmo.

UF	kmeans	Nome_Mesorregião	Área plantada_2020	Contagem_Municípios	Produtividade_Média_Ponderada	% CV Prod. Méd. Ponderado	Sinistralidade ponderada 2010 a 2020 (%)
PR	0.0	Centro Ocidental Paranaense	490.900	259	3.255,90	13.45	34.39
		Centro Oriental Paranaense	510.100	259	3.532,83	9.42	17.85
		Centro-Sul Paranaense	36.480	259	3.131,05	14.83	22.04
		Metropolitana de Curitiba	174.724	259	3.357,99	10.36	12.82
		Noroeste Paranaense	280.371	259	2.824,10	20.73	91.31
		Norte Central Paranaense	926.214	259	3.125,31	15.79	46.14
		Norte Pioneiro Paranaense	517.600	259	3.040,56	17.54	43.99
		Oeste Paranaense	240.251	259	3.258,21	22.91	54.91
		Sudeste Paranaense	345.600	259	3.276,00	8.24	14.56
	1.0	Centro Ocidental Paranaense	173.300	119	3.369,10	11.06	28.58
		Centro-Sul Paranaense	578.520	119	3.356,81	12.89	14.64
		Metropolitana de Curitiba	644	119	3.195,36	9.92	0.00
		Noroeste Paranaense	280	119	2.298,71	16.14	143.85
		Oeste Paranaense	778.043	119	3.328,47	18.90	49.82
		Sudeste Paranaense	27.400	119	3.099,88	8.31	13.17
		Sudoeste Paranaense	452.410	119	3.220,23	19.01	33.44

Figura 35: Tabela resumo K-Means por Mesorregião_PR

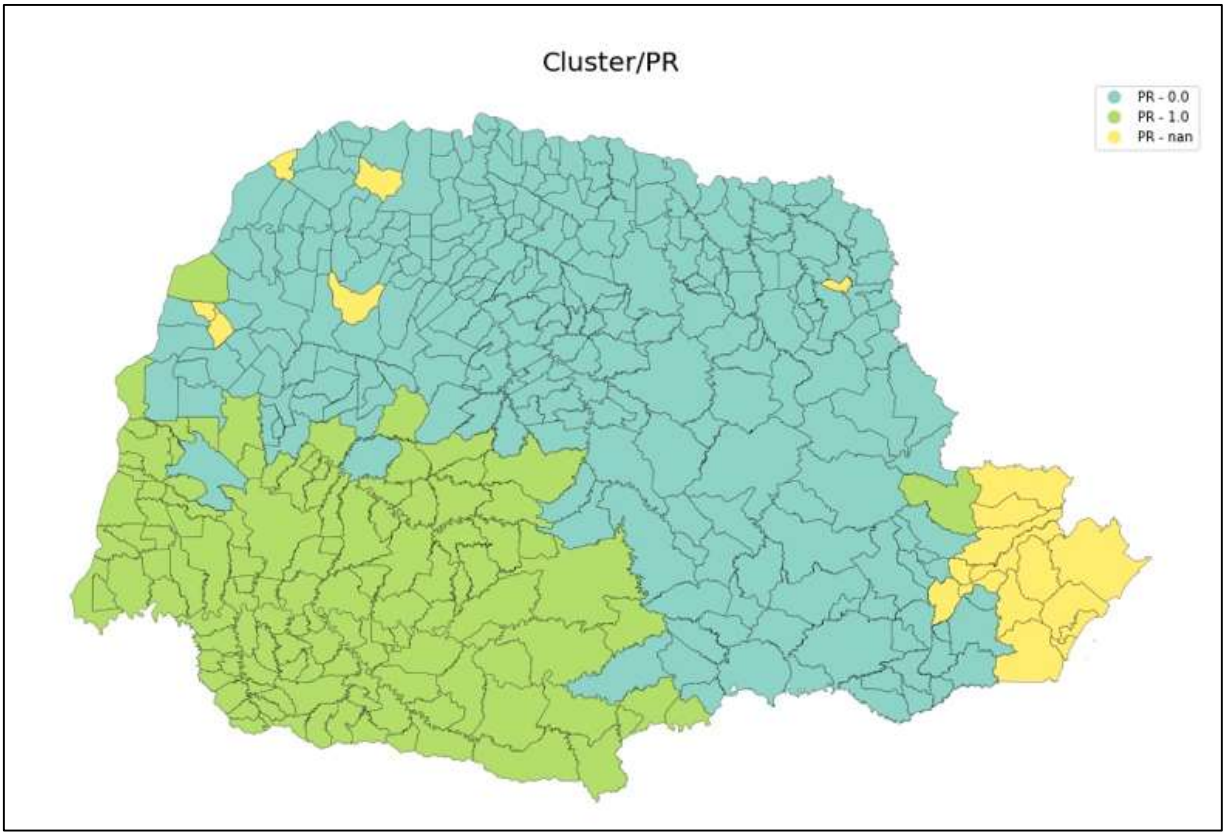

Figura 36: Mapa K-Means_PR

	UF	Geocódigo	Município (UF)	kmeans	Área plantada_2020	Média Prod 2010 a 2020
0	PR	4127502	Tibagi (PR)	0	100.000,00	3.570,18
1	PR	4119905	Ponta Grossa (PR)	0	70.500,00	3.610,00
2	PR	4127700	Toledo (PR)	0	70.400,00	3.233,64
3	PR	4104907	Castro (PR)	0	65.500,00	3.647,91
4	PR	4113700	Londrina (PR)	0	62.000,00	2.948,82
5	PR	4113205	Lapa (PR)	0	57.000,00	3.322,27
6	PR	4127403	Terra Roxa (PR)	0	54.000,00	3.029,27
7	PR	4117701	Palmeira (PR)	0	52.100,00	3.447,45
8	PR	4104303	Campo Mourão (PR)	0	48.500,00	3.323,09
9	PR	4113734	Luiziana (PR)	0	48.000,00	3.403,36

Figura 37: Rank dos principais municípios do cluster escolhido_PR

Rio Grande do Sul:

Através da plotagem dos clusteres no mapa do Rio Grande do Sul, o algoritmo captou a alta variabilidade que o sul do estado no geral tem, evidenciando a divisão dos clusteres basicamente entre norte e o sul do estado.

O cluster 1 apresenta uma concentração maior de área plantada combinado com uma produtividade média maior e um menor coeficiente de variação, o que torna nesta análise o cluster “ótimo”.

UF	kmeans	Nome_Mesorregião	Área plantada_2020	Contagem_Municipios	Produtividade_Média_Ponderada	% CV Prod. Méd. Ponderado	Sinistralidade ponderada 2010 a 2020 (%)
RS	0.0	Centro Ocidental Rio-grandense	287.645	123	2.414,83	24.06	55.48
		Centro Oriental Rio-grandense	265.605	123	2.532,94	24.90	74.32
		Metropolitana de Porto Alegre	147.823	123	2.475,96	20.68	75.15
		Noroeste Rio-grandense	50.300	123	2.684,90	33.46	71.03
		Sudeste Rio-grandense	503.383	123	2.158,13	28.61	63.68
		Sudoeste Rio-grandense	795.128	123	2.100,84	26.46	79.70
	1.0	Centro Ocidental Rio-grandense	536.790	280	2.723,48	26.10	41.28
		Centro Oriental Rio-grandense	42.160	280	2.929,84	23.49	59.89
		Nordeste Rio-grandense	314.314	280	3.167,50	17.93	31.87
		Noroeste Rio-grandense	3.033.139	280	2.874,30	25.93	39.17
		Sudoeste Rio-grandense	18.620	280	2.343,91	26.08	45.47

Figura 38: Tabela resumo K-Means por Mesorregião_RS

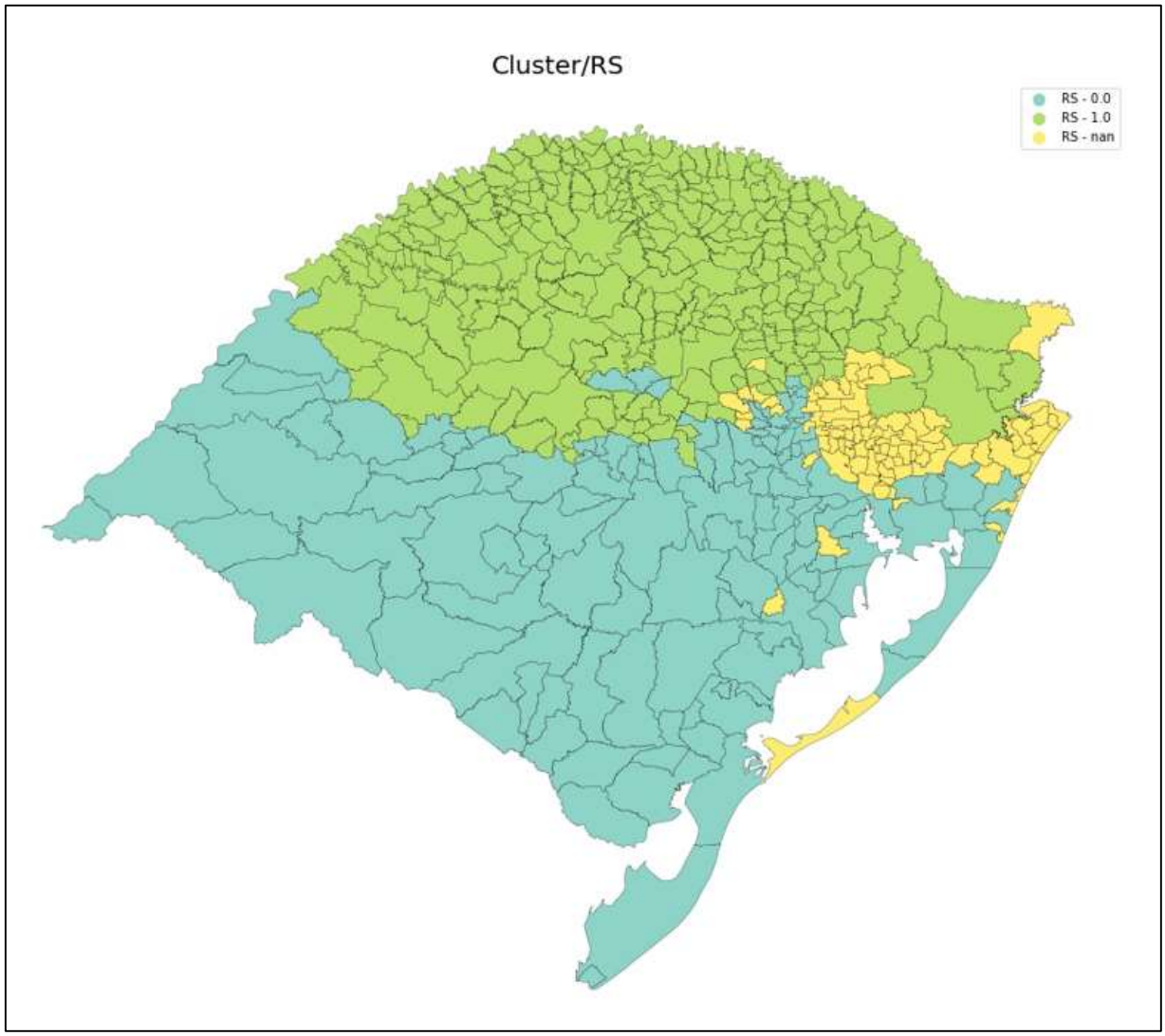

Figura 39: Mapa K-Means_RS

	UF	Geocódigo	Município (UF)	kmeans	Área plantada_2020	Média Prod 2010 a 2020
0	RS	4322202	Tupanciretã (RS)	1	149.100,00	2.797,64
1	RS	4313706	Palmeira das Missões (RS)	1	105.000,00	2.938,82
2	RS	4311205	Júlio de Castilhos (RS)	1	100.000,00	2.851,09
3	RS	4306106	Cruz Alta (RS)	1	92.000,00	2.823,09
4	RS	4311155	Jóia (RS)	1	81.000,00	2.400,36
5	RS	4316709	Santa Bárbara do Sul (RS)	1	76.000,00	3.233,45
6	RS	4318903	São Luiz Gonzaga (RS)	1	75.650,00	2.592,91
7	RS	4319158	São Miguel das Missões (RS)	1	74.000,00	2.566,55
8	RS	4304655	Capão do Cipó (RS)	1	63.700,00	2.363,09
9	RS	4309001	Giruá (RS)	1	58.400,00	2.675,00

Figura 40: Rank dos principais municípios do cluster escolhido_RS

Santa Catarina:

O cluster 0 deste estado foi o apresentou as três variáveis primárias que estamos avaliando para definição de cluster "ótimo" com resultado satisfatório, apesar da divisão em termos de área plantada ter sido bem equilibrada. O cluster 0 obteve no geral a maior área plantada, a maior produtividade média e o menor coeficiente de variação em termos de produtividade.

			Área plantada_2020	Contagem_Municípios	Produtividade_Média_Ponderada	% CV Prod. Méd. Ponderado	Sinistralidade ponderada 2010 a 2020 (%)
UF	kmeans	Nome_Mesorregião					
SC	0.0	Norte Catarinense	149.300	84	3.436,83	6.86	8.43
		Oeste Catarinense	30.050	84	3.170,90	11.99	21.46
		Serrana	165.136	84	3.146,52	17.13	24.57
		Sul Catarinense	4.150	84	2.989,08	18.44	104.08
		Vale do Itajaí	13.880	84	3.243,73	19.03	46.53
	1.0	Oeste Catarinense	300.157	89	3.110,19	18.35	15.09

Figura 41: Tabela resumo K-Means por Mesorregião_SC

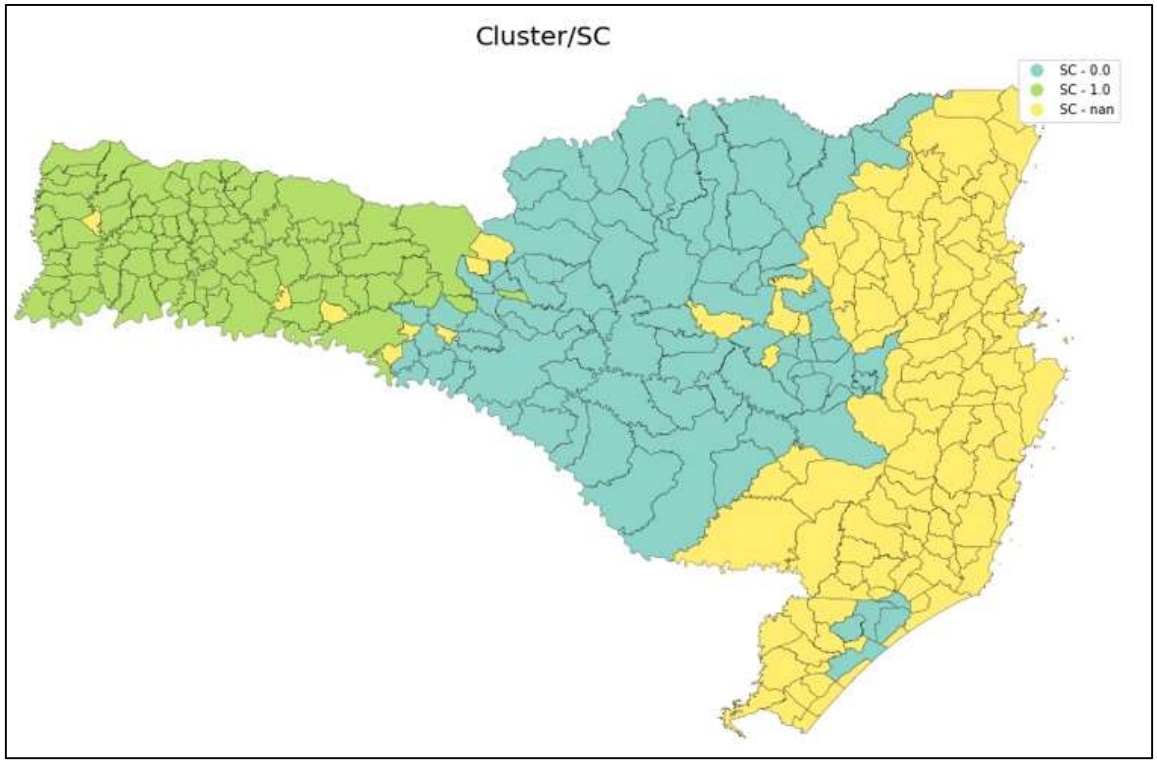

Figura 42: Mapa K-Means_SC

	UF	Geocódigo	Município (UF)	kmeans	Área plantada_2020	Média Prod 2010 a 2020
0	SC	4203600	Campos Novos (SC)	0	55.500,00	3.436,36
1	SC	4210100	Mafra (SC)	0	27.000,00	3.561,18
2	SC	4203808	Canoinhas (SC)	0	23.000,00	3.568,18
3	SC	4208104	Itaiópolis (SC)	0	18.000,00	3.362,73
4	SC	4212205	Papanduva (SC)	0	15.000,00	3.427,27
5	SC	4204806	Curitibanos (SC)	0	15.000,00	2.900,00
6	SC	4215000	Rio Negrinho (SC)	0	12.000,00	3.109,09
7	SC	4207908	Irineópolis (SC)	0	11.500,00	3.621,82
8	SC	4210308	Major Vieira (SC)	0	11.000,00	3.380,91
9	SC	4209300	Lages (SC)	0	10.000,00	3.234,55

Figura 43: Rank dos principais municípios do cluster escolhido_SC

São Paulo:

Para o estado de São Paulo, o cluster 0 agrupou os municípios de maior representatividade e maior produtividade, porém, a volatilidade em termos de produtividade não foi a menor das calculadas. Neste caso, um outro fator que corrobora também para o tornar o cluster “ótimo”, é o fato da concentração deste cluster estar na divisa do Paraná, próximo ao cluster “ótimo” avaliado para tal estado, agrupando de um certo modo regiões com comportamentos similares entre estados.

UF	kmeans	Nome_Mesorregião	Área plantada_2020	Contagem_Municípios	Produtividade_Média_Ponderada	% CV Prod. Méd. Ponderado	Sinistralidade ponderada 2010 a 2020 (%)
SP	0.0	Araçatuba	59.487	177	2.846,58	20.83	134.33
		Assis	288.716	177	2.949,35	15.04	40.46
		Bauru	92.542	177	3.261,27	16.07	31.53
		Campinas	1.350	177	3.105,56	23.96	23.04
		Itapetininga	297.440	177	3.383,62	19.93	17.24
		Macro Metropolitana Paulista	21.925	177	3.367,06	23.30	25.20
		Marília	8.900	177	2.603,45	12.88	23.32
		Piracicaba	1.700	177	2.353,81	7.80	87.53
		Presidente Prudente	62.054	177	2.703,42	18.20	108.18
		São José do Rio Preto	3.980	177	2.954,06	14.58	47.14
	1.0	Araraquara	19.018	187	2.860,03	15.89	33.32
		Araçatuba	441	187	2.423,00	30.55	0.00
		Bauru	16.961	187	2.996,61	14.84	40.20
		Campinas	40.410	187	2.968,69	14.52	16.64
		Macro Metropolitana Paulista	3.431	187	2.909,52	6.62	0.90
		Piracicaba	13.334	187	2.974,78	18.20	9.36
		Ribeirão Preto	140.819	187	2.914,08	15.78	54.81
		São José do Rio Preto	43.052	187	2.752,25	18.63	90.48
		Vale do Paraíba Paulista	2.770	187	3.108,26	21.06	0.00

Figura 44: Tabela resumo K-Means por Mesorregião_SP

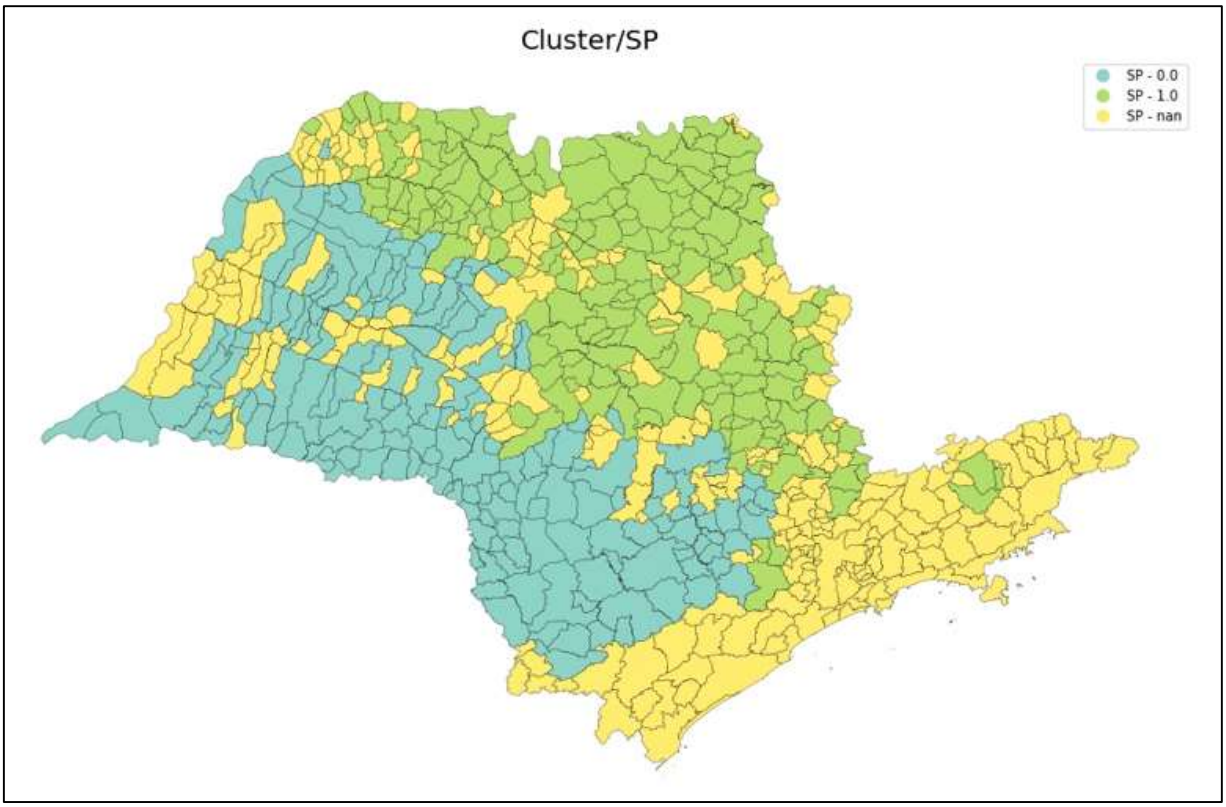

Figura 45: Mapa K-Means_SP

	UF	Geocódigo	Município (UF)	kmeans	Área plantada_2020	Média Prod 2010 a 2020
0	SP	3522406	Itapeva (SP)	0	81.000,00	3.665,18
1	SP	3521705	Itaberá (SP)	0	41.000,00	3.489,64
2	SP	3546405	Santa Cruz do Rio Pardo (SP)	0	31.000,00	3.077,18
3	SP	3535309	Palmital (SP)	0	28.750,00	2.874,91
4	SP	3528809	Maracaí (SP)	0	26.935,00	2.893,27
5	SP	3510005	Cândido Mota (SP)	0	25.600,00	2.961,00
6	SP	3522307	Itapetininga (SP)	0	22.500,00	2.982,18
7	SP	3523206	Itararé (SP)	0	22.000,00	3.456,82
8	SP	3542206	Rancharia (SP)	0	21.500,00	2.797,73
9	SP	3535804	Paranapanema (SP)	0	21.000,00	3.317,64

Figura 46: Rank dos principais municípios do cluster escolhido_SP

Tocantins:

O estado do Tocantins foi o único que agrupou os municípios em três classes. Considerando a concentração de área plantada e maior produtividade, o cluster 1 atende esses requisitos, apesar da volatilidade de produtividade ser maior que os outros dois. Atrelado ao fato de concentrar os municípios mais representativos, o cluster 1 foi sugerido como "ótimo" com ressalvas. Ficando a recomendação no caso de um maior apetite de risco por essa região em função da classificação como "cluster ótimo", a oferta de produtos e precificação adequada considerando as particularidades em termos de volatilidade.

UF	kmeans	Nome_Mesorregião	Área plantada_2020	Contagem_Municípios	Produtividade_Média_Ponderada	% CV Prod. Méd. Ponderado	Sinistralidade ponderada 2010 a 2020 (%)
TO	0.0	Ocidental do Tocantins	248.972	37	2.845,86	9.62	41.43
		Oriental do Tocantins	149.258	37	2.851,56	11.01	33.02
	1.0	Ocidental do Tocantins	220.947	30	2.880,84	15.05	36.93
		Oriental do Tocantins	208.609	30	2.881,09	10.10	41.45
	2.0	Ocidental do Tocantins	38.490	14	2.588,21	10.52	111.05
		Oriental do Tocantins	91.533	14	3.054,62	5.90	30.31

Figura 47: Tabela resumo K-Means por Mesorregião_TO

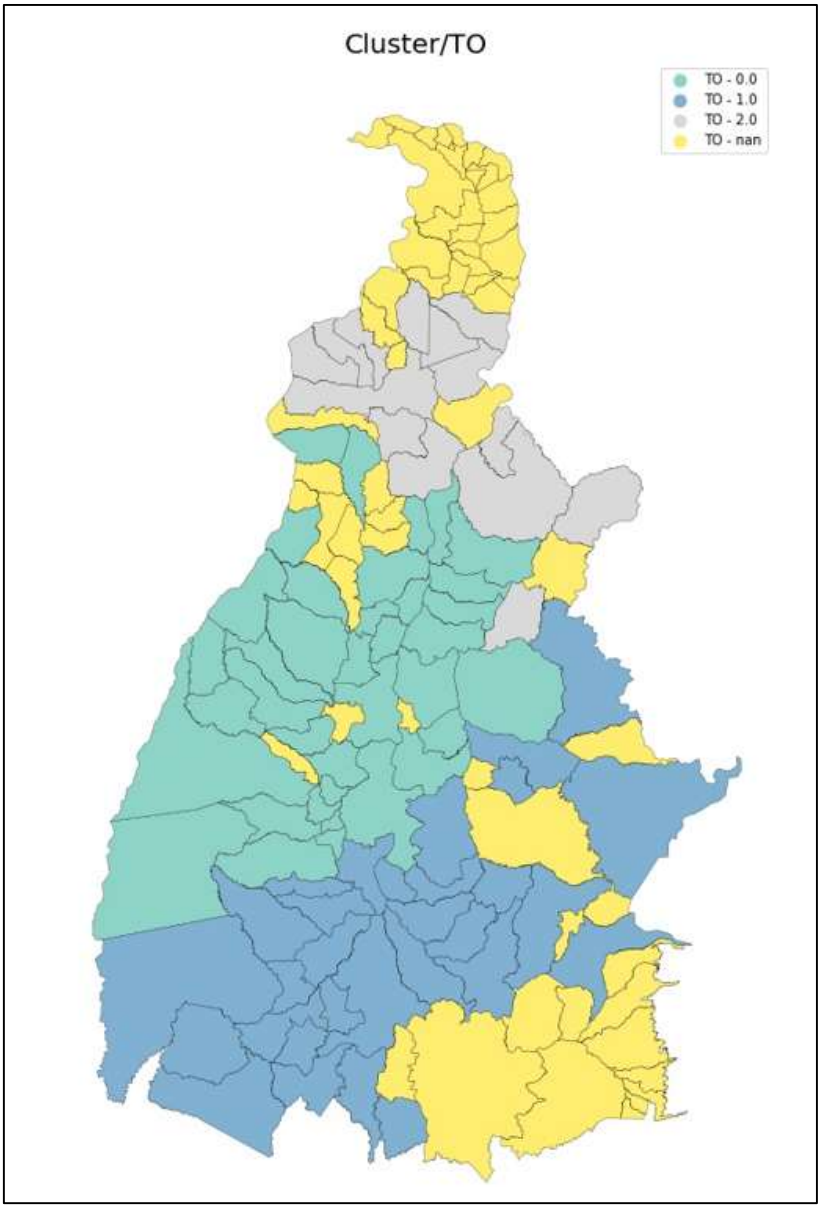

Figura 48: Mapa K-Means_TO

	UF	Geocódigo	Município (UF)	kmeans	Área plantada_2020	Média Prod 2010 a 2020
0	TO	1716604	Peixe (TO)	1	57.194,00	2.801,36
1	TO	1712702	Mateiros (TO)	1	46.000,00	2.890,55
2	TO	1718907	Santa Rosa do Tocantins (TO)	1	37.000,00	2.906,18
3	TO	1713601	Monte do Carmo (TO)	1	36.000,00	2.795,45
4	TO	1700707	Alvorada (TO)	1	28.181,00	2.865,18
5	TO	1720655	Silvanópolis (TO)	1	24.000,00	2.793,55
6	TO	1703701	Brejinho de Nazaré (TO)	1	23.500,00	2.890,91
7	TO	1707652	Figueirópolis (TO)	1	19.417,00	2.902,64
8	TO	1720853	Sucupira (TO)	1	15.479,00	2.989,91
9	TO	1703867	Cariri do Tocantins (TO)	1	14.963,00	2.963,00

Figura 49: Rank dos principais municípios do cluster escolhido_TO

Cabe ressaltar que esta análise não tem por objetivo recomendar que uma seguradora não oferte produtos para uma determinada região, mas sim tenha indícios de locais que possuem similaridade em termos de produtividade, área, coeficiente de variação e comportamento precipitação dos últimos anos, considerando precipitação acumulada média e coeficiente de variação média por safra dos últimos 30 anos, em conjunto com o histórico de contratação de seguro agrícola através do PSR, e com isso auxilie na dispersão geográfica e como consequência oferta de produtos aderentes a realidade de cada região.

Evidentemente os municípios com maiores indicadores de variabilidade são mais arriscados que outros que possuem comportamento mais homogêneo e demandam mais atenção. A ilustração abaixo demonstra os municípios que tenderiam a uma homogeneidade no comportamento dentro de cada estado com base nos clusteres selecionados como “ótimos” após avaliação desta ferramenta.

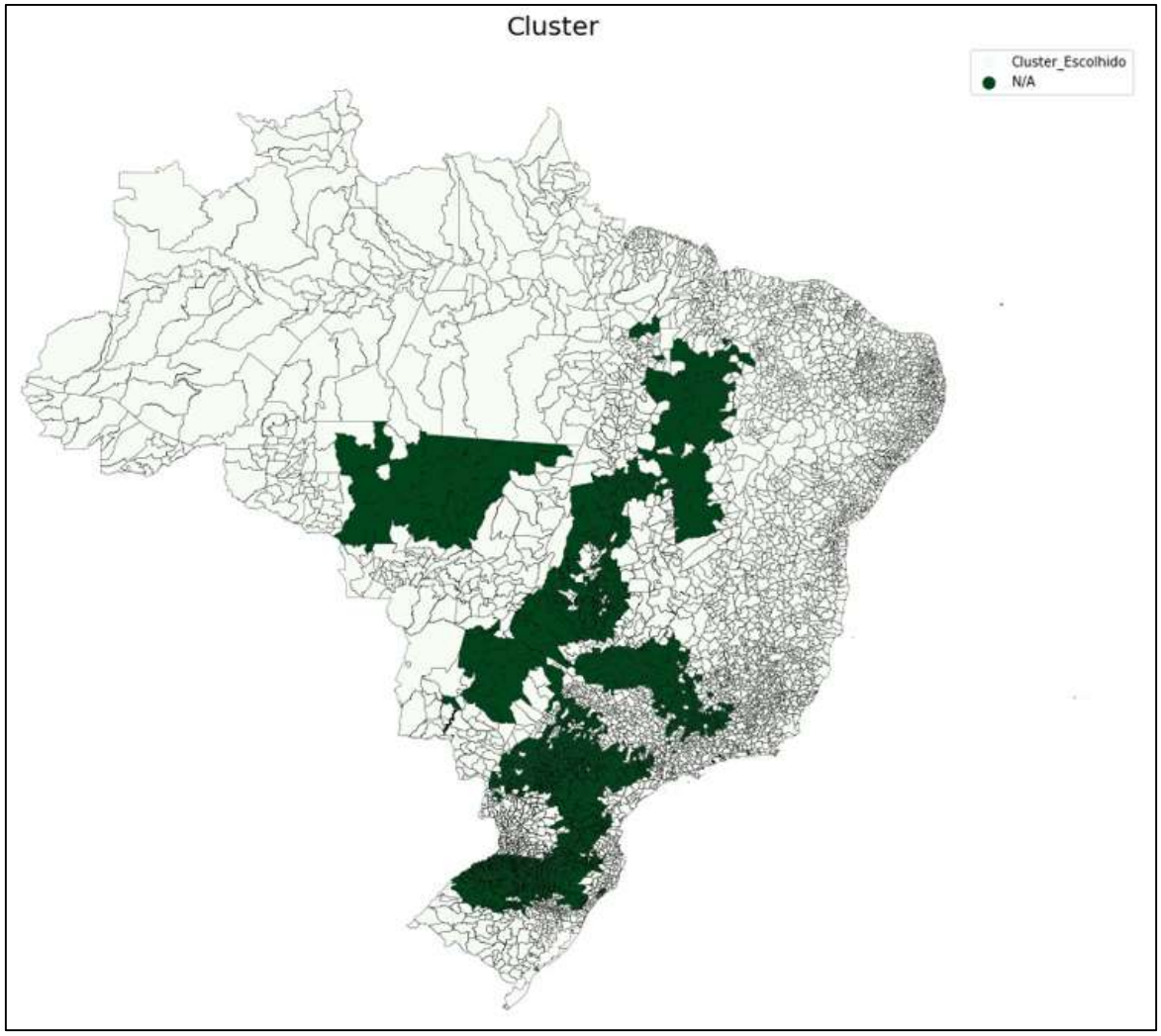

Figura 50: Cluster por estado - Mapa do Brasil

6 Conclusão

Como dito anteriormente, entendemos que a ciência de dados não tem como objetivo substituir a análise humana e desconsiderar experiências e conhecimentos das pessoas envolvidas na tomada de decisão. Porém, existem ferramentas excelentes e que podem oferecer insights de modo a contribuir para que a solução de uma determinada questão seja atingida com uma maior satisfação. Além de proporcionar uma maior agilidade e possibilidade de extrair indicadores enriquecedores para o negócio.

Visto isso, através desse estudo, foi possível avaliar as subdivisões que os estados têm em relação a produção de soja no Brasil, no qual foi considerado a produtividade média de 2010 a 2020; coeficiente de variação desta produtividade média, que traduz o risco embutido; área plantada, dando maior relevância a municípios que possuem a atividade agrícola desta cultura mais desenvolvida; informações relativas a contratação de seguro através dos dados do PSR, área segurada 2021 e sinistralidade de 2010 a 2021, esta última também com o intuito de indicar o grau de risco associado a operação; e as médias de precipitação acumulada e suas volatilidades dos últimos 30 anos separados pelos meses que compõe a safra de verão de uma forma geral no país (setembro, outubro, novembro, dezembro, janeiro e março). Buscando dessa forma similaridade entre tais variáveis e possibilitando uma avaliação de dispersão de risco aderente a volatilidade de cada região, adequação das condições ofertadas, como coberturas, taxas e clausulados que estejam em linha com a região que se deseja atuar.

Na análise deste estudo se identificou grupos de municípios similares por estado que tendem a ter um desempenho ou potencial produtivo maior e/ou menos volátil. De todo modo, permite, caso seja de interesse das Seguradoras, avaliar locais com maior risco embutindo, propondo condições mais adequadas, que no geral tendem a resultar em taxas mais agravadas por conta do alto risco relativo, ou mesmo visando a elaboração de um produto novo que atenda regiões mais complexas a fim de que proteja a carteira da companhia e buscando também um lado social através do enfoque de segurança alimentar.

7 Referências bibliográficas

CANAL RURAL; **El Niño ou La Niña: qual o fenômeno mais prejudicial para o agro brasileiro?** disponível em: https://www.canalrural.com.br/noticias/tempo/el-nino-ou-la-nina/, 2020.

Cassiano, Kelia Mara; **Análise de Séries Temporais Usando Análise Espectral Singular (SSA) e Clusterização de Suas Componentes Baseada em Densidade**; Tese (Doutorado em Engenharia Elétrica) - Departamento de Engenharia Elétrica - PUC-Rio, Rio de Janeiro, 2014.

CEPEA/ ESALQ-USP; PIB-AGRO/CEPEA: **PIB do agro cresce 8,36% em 2021; participação no PIB brasileiro chega a 27,4%**; disponível em: https://www.cepea.esalq.usp.br/br/releases/pib-agro-cepea-pib-do-agro-cresce-8-36-em-2021-participacao-no-pib-brasileiro-chega-a-27-4.aspx, 2022.

Cnseg; **Governo divulga portaria sobre Seguro Rural**; disponível em: https://cnseg.org.br/noticias/governo-divulga-portaria-sobre-seguro-rural.html, 2022.

CONAB - Companhia Nacional De Abastecimento; **Acompanhamento da Safra Brasileira**; disponível em: https://www.conab.gov.br/info-agro/safras, 2022.

Costi, Guilherme; **LAMBDA3**; disponível em: https://www.lambda3.com.br/2020/03/aprendizagem-nao-supervisionada/, 2020.

Dendroid; **KMeans Clustering algorithm explained**; disponível em: K-means Clustering algorithm explained - dendroid, 2022.

Education Ecosystem; **Understanding K-means Clustering in Machine Learning**; disponível em: https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1, 2018.

EMBRAPA - Empresa Brasileira de Pesquisa Agropecuária; **Soja em Números**; disponível em: https://www.embrapa.br/soja/cultivos/soja1/dados-economicos, 2022.

Escovedo, Tatiana; **Machine Learning: Conceitos e Modelos - Parte II: Aprendizado Não-Supervbisionado**; disponível em: https://tatianaesc.medium.com/machine-learning-conceitos-e-modelos-parte-ii-aprendizado-n%C3%A3o-supervisionado-fb6d83e4a520, 2020.

FIA BUSINESS SCHOOL; **Agronegócio: o que é, como funciona e setores**; disponível em: https://fia.com.br/blog/agronegocio/, 2021.

Géron, Aurélien; **Mãos à obra: aprendizado de máquina com Scikit-Learn & TensorFlow**: Conceitos, Ferramentas e Técnicas para a Construção de Sistemas Inteligentes. ALTA BOOKS, 2019.

Ministério da Agricultura, Pecuária e Abastecimento; **Raio X do PSR - Relatório 2021**; disponível em: https://www.gov.br/agricultura/pt-br/assuntos/riscos-seguro/seguro-rural/dados/relatorios/relatorio-geral-psr-2021-final.pdf, 2022.

Provost, Foster; Fawcett, Tom; **Data Science para Negócios: O que Você Precisa Saber Sobre Mineração de Dados e pensamento Analítico de Dados**; ALTA BOOKS, 2016.

SIDRA - Sistema IBGE de Recuperação Automática; **Produção Agrícola Municipal - PAM 2020**; disponível em: https://sidra.ibge.gov.br/pesquisa/pam/tabelas, 2022.

Silva, Emerson R. M.; Barbosa; Ivan C. C., Silv., Helder J.F; Costa, Luiz G.S.; Rocha, Edson J.P.; **Análise do Desempenho da Estimativa de Precipitação do Produto CHIRPS para a Sub-Bacia do Rio Apeú; Revista Brasileira de Geografia Física**, v.13, p 1094-1105,2020.

SUSEP - Superintendência de Seguros Privados; **Seguro Rural**; disponível em: http://www.susep.gov.br/menu/informacoes-ao-publico/planos-e-produtos/seguros/seguro-rural, 2022.

University of California, Santa Barbara; **CHIRPS: Rainfall Estimates from Rain Gauge and Satellite Observations**; disponível em: https://www.chc.ucsb.edu/data/chirps, 2022.

Wikipedia; **Silhouette (clustering);** disponível em: https://en.wikipedia.org/wiki/Silhouette_(clustering), 2022.