

O impacto climático sobre a produtividade do milho americana: uma perspectiva através de modelos multiníveis

Gustavo Ramalho

24 de junho de 2024

Estudo de caso da disciplina de Modelagem Estatística, submetido como trabalho final da matéria válida pelo curso de Ciência de Dados & Inteligência Artificial da EMAP.
Professor: Luiz Max Carvalho

Resumo

O objetivo deste trabalho é modelar a produtividade (produção por área) do milho nos principais estados americanos através de dados climáticos. Para isso, utilizaremos modelos de regressão multiníveis que serão capazes de captar características inerentes ao nosso contexto, produzindo estimativas melhores do que modelos mais simples. Dessa forma, buscaremos responder uma série de questões propostas e entender o contexto de nosso problema principal, e isso será feito através de etapas de entendimento dos dados por meio de análise exploratória, e diferentes testes com os modelos propostos. Por fim, as conclusões mostram que o fator temporal se revelou como ímpar para as estimativas, e variáveis climáticas como diferença de temperatura e precipitação em meses específicos possuem impacto significativo na produtividade final.

Palavras-chave: Produtividade, Modelos de Regressão Multiníveis, Dados Climáticos, Modelagem Estatística

Contents

1	Introdução	3
2	Revisão de literatura	4
3	Metodologia	4
3.1	Tratamento dos dados	4
3.1.1	Dados Climáticos (NOAA)	4
3.1.2	Dados de produção (USDA)	6
3.2	Modelagem	7
3.2.1	Modelo de Regressão Linear Múltipla	7
3.2.2	Modelos Multiníveis com Intercepto Variável	7
3.2.3	Modelos Multiníveis com Intercepto e Coeficientes Variáveis	8
3.3	Métodos numéricos	8
3.3.1	Mínimos Quadrados	8
3.3.2	Máxima Verossimilhança Restrita	8
3.4	Crítérios de avaliação	9
4	Resultados	10
4.1	Análise Exploratória	10
4.2	Modelos ajustados	13
4.2.1	Regressão linear múltipla simples	14
4.2.2	Modelos multiníveis com intercepto variáveis	14
4.2.3	Modelos Multiníveis com Intercepto e Coeficientes Variáveis	18
4.2.4	Outras abordagens e resultados	19
5	Conclusão	20


1 Introdução

Os Estados Unidos destacam-se como o principal produtor de milho mundial, um cultivo fundamental para a economia agrícola e para a cadeia de suprimentos alimentar e energética do país. O milho serve a uma gama diversificada de usos, desde ração animal até a fabricação de etanol, tornando-se essencial para o planejamento agrícola, as políticas de subsídios e as estratégias de mercado. A capacidade de prever com precisão a produtividade (produção por área cultivada) do milho é, portanto, de suma importância para vários setores da economia.

As variações nessa produtividade, ou *yield*, são profundamente impactadas pelas condições climáticas. Elementos como temperatura e precipitação podem alterar significativamente os níveis de produção, exigindo análises detalhadas para uma compreensão efetiva de tais efeitos. Neste cenário, a análise que foca na previsão do *yield* pode revelar como as mudanças climáticas afetam a produtividade do milho, fornecendo *insights* essenciais para adaptações agrícolas.

Nesse sentido, é crucial reconhecer as complexidades inerentes à nossa análise devido à vasta extensão territorial dos Estados Unidos, além do claro avanço tecnológico em técnicas de cultivo ao longo dos anos. Seria simplista assumir uniformidade no cultivo do milho em todos os estados considerados neste estudo? Ou seja, diferentes climas regionais podem influenciar significativamente o *yield*? Além disso, os avanços anuais - não explicados por mudanças climáticas - são relevantes para a análise? Por fim, quais fases do plantio são mais importantes para determinar a produção final?

Para responder a essas perguntas, este trabalho propõe desenvolver modelos de regressão linear multiníveis para prever as variações no *yield* de milho em 18 dos principais estados produtores nos Estados Unidos, que conforme indicado pelo *United States Department of Agriculture* (USDA), são responsáveis por mais de 90% da produção nacional anual. Ao utilizar dados climáticos como base para nossas estimativas, buscamos identificar padrões e correlações que possam explicar a produtividade devido a mudanças climáticas. Esta abordagem não só poderá enfatizar as possíveis diferentes condições agrícolas estaduais, mas também destaca os períodos mais vulneráveis do ciclo de cultivo. Portanto, será necessário elencarmos quais covariáveis vão explicar melhor as variações no *yield*, ou seja, as combinações de período (mês) e variável climática presentes em nosso *dataset*.

Para este estudo, utilizarei exclusivamente dados climáticos coletados e divulgados diariamente pelo *National Oceanic and Atmospheric Administration* (NOAA). Estes dados abrangem variáveis climáticas essenciais como temperatura, precipitação e luminosidade. Além disso, será integrado informações sobre produção de milho fornecidas pelo USDA, o que possibilitará acessar diretamente a variável resposta, o *yield* de milho. Os códigos utilizados para extração e tratamento dos dados foram feitos em Python, assim como as visualizações. O ajuste dos modelos foi realizado através do R pela maior simplicidade. Cada uma dessas etapas pode ser encontrada neste repositório do [GitHub](#) .

Em um primeiro momento, será realizada uma revisão de literatura para entender fatores que influenciam a produtividade e identificar metodologias previamente utilizadas. Após isso, passaremos por uma etapa de tratamento dos dados, onde será possível se familiarizar com as variáveis preditoras e variável resposta. Em seguida, abordarei nossas metodologias, como modelos a serem utilizados e métodos de avaliação. Por fim, discutirei nossos resultados, indo da análise exploratória até a criação dos modelos propriamente ditos.

2 Revisão de literatura

O plantio do milho tem características particulares que precisam ser analisadas antes de começarmos a lidar efetivamente com os dados. Nesse sentido, algumas dúvidas surgem em relação ao nosso tema: quais variáveis climáticas são essencialmente importantes? Qual o período de plantio desse cereal?

Uma pesquisa conduzida por [Joshi et al. \[2021\]](#) demonstra que a utilização de dados meteorológicos, como precipitação total, temperatura média do ar, e a diferença entre as temperaturas máxima e mínima, pode fornecer estimativas confiáveis dos rendimentos de milho no Cinturão Central de Milho dos EUA. A pesquisa avalia o efeito dessas variáveis em diferentes escalas temporais — semanal, quinzenal e mensal — revelando que a temperatura média e a diferença de temperatura durante os meses críticos de julho e agosto, além da precipitação em junho e julho, são indicadores significativos para o rendimento do milho. Apesar do cinturão do milho abranger um pequeno grupo de estados no centro-norte do país, as informações podem ser significativas para nossa análise que se propõe mais abrangente.

Além disso, a inclusão de T_{diff} nos modelos mostrou melhorias consideráveis nas estimativas de rendimento. Este achado é crucial, pois indica que não apenas as temperaturas médias, mas também as flutuações diárias de temperatura podem ter efeitos substanciais nos rendimentos das culturas. É interessante notar que podemos criar essa variável, dado que temos informações de temperatura máxima e mínima.

Outra questão importante é compreender a cronologia das fases de crescimento do milho. Com isso em mente, conseguiremos ter uma melhor noção de fases mais críticas e suscetíveis a oscilações na temperatura, que podem afetar a produção final. O plantio do milho geralmente começa em maio, quando as temperaturas do solo alcançam condições adequadas para a germinação das sementes, idealmente em torno de 10°C (50°F) [Shaw and Newman \[1985\]](#).

Após o plantio, há um intenso período de crescimento vegetativo que se estende até junho. Em, julho, o milho inicia a transição para a fase reprodutiva e estresses como falta de água e altas temperaturas podem afetar o potencial reprodutivo. Por fim, em agosto, ocorre a fase final de polinização e maturação, que são também suscetíveis a mudanças climáticas e hídricas [Shaw and Newman \[1985\]](#).

Com essas informações, temos um direcionamento melhor sobre quais variáveis climáticas e quais meses podemos dar maior atenção, já que a combinação desses dois fatores vão compor nossas variáveis preditivas. Posteriormente, em nossa análise exploratória, buscaremos também ver como é o comportamento de cultivo separadamente em cada estado, e como a temperatura vai se comportar em cada um deles.

3 Metodologia

3.1 Tratamento dos dados

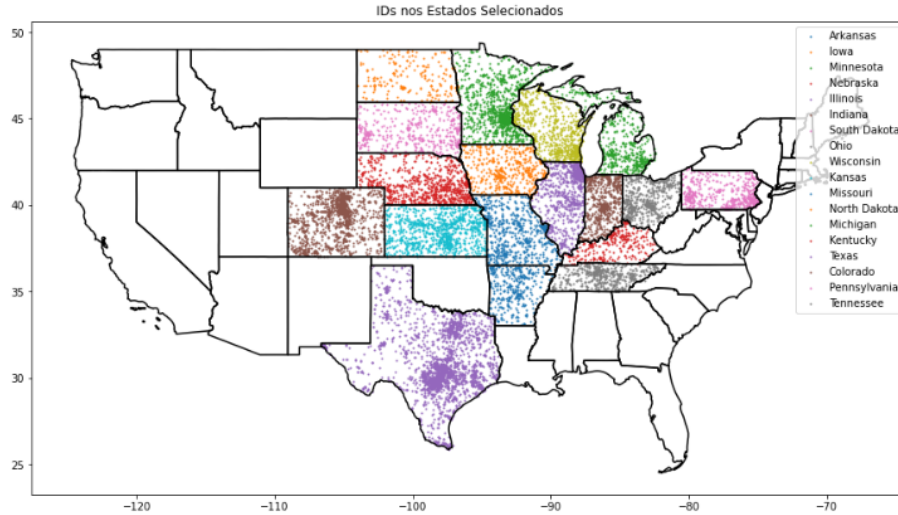
3.1.1 Dados Climáticos (NOAA)

Antes de chegarmos no *dataset* final pronto para modelagem, há um processo considerável de manipulação e tratamento dos dados, principalmente com os dados de clima.

As informações disponibilizadas pelo NOAA ([disponíveis aqui](#)) são basicamente dados de variáveis climáticas ao redor do mundo, em que cada arquivo *csv* contém as informações

de um ano específico. Essas informações são dadas por sensores no solo com latitude e longitude, e a imagem abaixo mostra como eles estão distribuídos nos estados de nosso interesse:

Figure 1: Sensores nos 18 estados de interesse



Cada um desses pontos são sensores que enviam dados climáticos de diversos tipos diariamente da localização em que estão. Iremos realizar uma filtragem para selecionar apenas os sensores mostrados no mapa, que fazem parte dos 18 estados principais, que irão compor esse trabalho. Estes são:

- Arkansas (AR)
- Colorado (CO)
- Iowa (IA)
- Illinois (IL)
- Indiana (IN)
- Kansas (KS)
- Kentucky (KY)
- Michigan (MI)
- Minnesota (MN)
- Missouri (MO)
- North Dakota (ND)
- Nebraska (NE)
- Ohio (OH)
- Pennsylvania (PA)
- South Dakota (SD)
- Tennessee (TN)
- Texas (TX)
- Wisconsin (WI)

Após essa etapa, temos basicamente os sensores com seus *reports* diários das diversas variáveis climáticas, e precisamos realizar 2 processos:

- Filtrar quais anos iremos usar em nossa modelagem. Iremos considerar a partir de 1973 até 2023, que são anos que temos uma quantidade interessante de dados.
- Filtrar quais variáveis climáticas usaremos. Seguindo as informações da literatura, vamos focar em temperatura máxima (TMAX), temperatura mínima (TMIN), temperatura média (TMEAN), diferença de temperaturas (TDIFF) e precipitação (PRCP).

Agora que temos os dados filtrados com os anos que usarei e as variáveis, é necessário calcular a média dos valores dos sensores para finalmente termos nosso *dataset* final.

Como visto na literatura que é razoável considerarmos as variações climáticas mensalmente, irei coletar a média mensal de cada uma das variáveis de interesse nos meses de plantio (5 a 8). Assim, temos nossas variáveis preditivas:

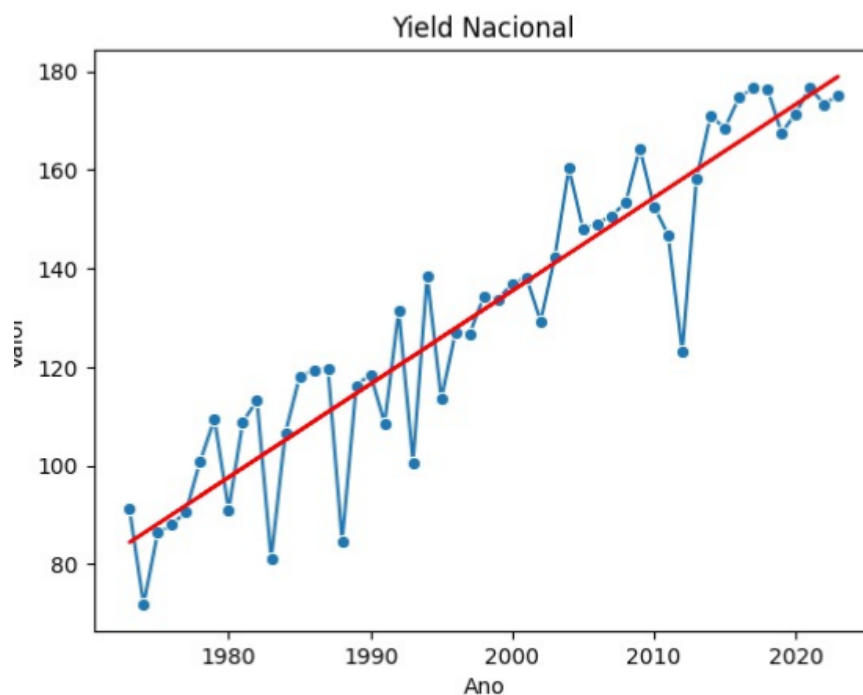
- State;
- Year;
- TMAX_5, TMAX_6, TMAX_7, TMAX_8;
- TMIN_5, TMIN_6, TMIN_7, TMIN_8;
- TDIFF_5, TDIFF_6, TDIFF_7, TDIFF_8;
- TMEAN_5, TMEAN_6, TMEAN_7, TMEAN_8;
- PRCP_5, PRCP_6, PRCP_7, PRCP_8.

É necessário ressaltar que, apesar da grande quantidade de covariáveis, não tenho a intenção de utilizar todas elas ao mesmo tempo - não queremos complexidade exagerada ao nosso modelo, muito menos multicolinearidade. Por isso, na etapa de modelagem, teremos um processo para realização de escolhas e testes com diferentes covariáveis.

3.1.2 Dados de produção (USDA)

A nossa variável resposta, ou seja, o *yield* do milho, será dada pelo USDA ([disponível aqui](#)) e será mais simples de conseguir do que as variáveis preditoras. Temos as informações de produtividade dadas pelo próprio órgão, e basta selecionar para nossos estados de interesse. A título de curiosidade e também para comparações futuras no trabalho, o comportamento dessa produtividade segue aumentando ano após ano, e a tendência revela uma "melhor utilização" das terras produzidas.

Figure 2: Tendência do *yield*



3.2 Modelagem

A análise da produção agrícola, especialmente em culturas significativas como o milho, exige uma abordagem estatística que possa capturar complexidades tanto em nível local quanto mais amplo. A escolha de modelos multiníveis para este estudo reflete nossa necessidade de abordar a heterogeneidade inerente aos dados agrícolas, que são influenciados por uma variedade de fatores climáticos, geográficos e temporais.

Os modelos multiníveis são particularmente adequados para este tipo de análise porque permitem que diferenciamos não apenas entre variações individuais na produção de milho em diferentes estados, mas também ao longo de diferentes anos. Será possível considerar múltiplos níveis de agrupamento nos dados, como estados e anos, e ajustarei a influência de preditores dentro de cada grupo. Isso quer dizer que é possível capturar nuances inerentes a estados específicos, como o efeito da localização e características geográficas únicas, bem como também é possível capturar mudanças ao longo do tempo que podem influenciar a produção devido a fatores como melhores práticas agrícolas.

No desenvolvimento desta análise, aplicarei uma regressão linear simples e duas principais configurações de modelos multiníveis: com variação apenas no intercepto e com variação no intercepto e coeficientes, cada um trazendo uma lente distinta sobre os dados para identificar a melhor abordagem para entender as características da produção de milho.

3.2.1 Modelo de Regressão Linear Múltipla

Inicialmente, testarei um modelo de regressão linear múltipla para identificar as relações básicas entre as variáveis climáticas e a produção de milho. Este passo inicial nos ajudará a compreender as associações diretas e servirá como ponto de partida para análises mais complexas. A equação do modelo é dada por:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \epsilon$$

Onde \hat{y} é o valor previsto, β_0 é o intercepto e $\beta_1, \beta_2, \dots, \beta_k$ são coeficientes para cada uma das k variáveis preditoras. O processo para estimar esses coeficientes e intercepto será dado pelo *OLS*, que vamos explicar na próxima sessão.

3.2.2 Modelos Multiníveis com Intercepto Variável

Em um segundo momento, aplicarei modelos multiníveis onde apenas o intercepto varia entre os grupos (estados ou anos). Esta abordagem permitirá avaliar se a produção base de milho difere significativamente entre os estados e ao longo dos anos, assumindo que a resposta ao clima é uniforme em todos os grupos. Esta análise fornecerá *insights* sobre as variações de base na produção de milho que podem ser atribuídas a características intrínsecas de cada estado ou às condições anuais. A fórmula desse modelo é dada por:

$$\hat{y} = \alpha_{j[i]} + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki} + \epsilon_i$$

onde \hat{y} é o valor previsto, $\alpha_{j[i]}$ é o intercepto variável para o grupo j ao qual a i -ésima observação pertence. É o termo que permite o intercepto variar entre grupos. β_1, \dots, β_k são coeficientes para cada k variável independente.

3.2.3 Modelos Multiníveis com Intercepto e Coeficientes Variáveis

Nesta etapa, exploraremos modelos onde tanto o intercepto quanto coeficientes de inclinação para as variáveis relacionadas ao ano são variáveis. Esta abordagem mais complexa permitirá que o modelo acomode como diferentes estados ou épocas reagem às variações climáticas. Por exemplo, a produção de milho nos estados pode ser mais afetada por mudanças específicas de temperatura de um ano para outro, como uma alteração na precipitação atualmente pode ser menos significativa do que alterações em outros períodos.

Como teremos coeficientes fixos e também aleatórios, a fórmula geral será dada por:

$$y = X\beta + Z\gamma + \epsilon$$

Onde:

- X, β : Matriz de desenho para os efeitos fixos e vetor dos coeficientes dos efeitos fixos, respectivamente;
- Z, γ : Matriz de desenho para os efeitos aleatórios e vetor para os coeficientes aleatórios que segue uma distribuição $N(0, G)$, respectivamente;
- ϵ : vetor dos erros residuais que segue uma distribuição $N(0, R)$.

É válido ressaltar que também podemos representar o último caso com essa fórmula (apenas o intercepto aleatório), já que conseguimos considerar o intercepto aleatório como um vetor no lugar da matriz Z .

3.3 Métodos numéricos

3.3.1 Mínimos Quadrados

No momento em que criarmos o modelo de regressão linear, iremos utilizar o **Método dos Mínimos Quadrados (OLS)**. Temos a intenção de fornecer um resultado inicial mais simples, e uma consequente fácil interpretação também. O método dos mínimos quadrados nos fornece isso e com ele conseguiremos ter uma direção melhor sobre nossos dados. O objetivo do método é minimizar a soma dos quadrados das diferenças entre os valores observados e os valores estimados pelo modelo. Para isso, estima-se os valores dos coeficientes do modelo que minimizam essa diferença:

$$S = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Calcula-se as derivadas parciais do intercepto e dos coeficientes, igualando cada um a 0 e resolvendo a equação resultante.

3.3.2 Máxima Verossimilhança Restrita

Para as modelagens multiníveis, temos a intenção de utilizar a **máxima verossimilhança restrita (REML)**.

Em um modelo linear misto, temos a seguinte fórmula, já comentada anteriormente:

$$y = X\beta + Z\gamma + \epsilon$$

Como também foi dito, γ é o vetor de efeitos aleatórios que segue a distribuição $N(0, G)$ e ϵ é o vetor dos erros que segue a distribuição $N(0, R)$. Se usássemos Máxima Verossimilhança, estimaria β, G, R maximizando a função de verossimilhança, que é baseada na distribuição conjunta dos dados. O problema é que o método considera todos os dados como parâmetros fixos, incluindo os efeitos aleatórios. Isso nos levaria a estimativas enviesadas dos componentes de variância dos efeitos aleatórios, porque o *ML* tende a subestimar a variância desses efeitos. Nesse sentido, o *REML* é usado para ajustar modelos multiníveis porque ele ajuda a obter estimativas mais precisas dos parâmetros de variância dos componentes aleatórios do modelo. Isso é alcançado ajustando os parâmetros de modo que a verossimilhança dos resíduos seja maximizada, após termos removido os efeitos dos preditores fixos.

O método consegue fazer isso projetando y para um espaço ortogonal X usando uma matriz de projeção, considerando o número de parâmetros fixos. Assim, temos estimativas para G e R menos enviesadas, removendo o efeito dos parâmetros fixos antes de estimarmos a variância.

3.4 Critérios de avaliação

Para garantir a precisão e a eficácia dos modelos estatísticos desenvolvidos na análise, adotaremos uma série de critérios de avaliação. Utilizaremos o MSE (*Mean Square Error*), o coeficiente de determinação (R^2), o AIC (*Akaike Information Criterion*) e BIC (*Bayes Information Criterion*).

O R^2 é uma medida estatística que indica a proporção da variância na variável dependente que é explicada pelas variáveis independentes no modelo. Um valor de R^2 próximo de 1 sugere que o modelo tem uma capacidade excelente de prever a variável dependente, enquanto um valor próximo de 0 indica que o modelo não explica adequadamente a variabilidade dos dados.

O MSE mede a média dos quadrados dos erros entre os valores preditos pelo modelo e os valores reais, o que nos dará a informação do quão longe nossos resultados estão dos valores observados.

O AIC é um método amplamente utilizado para seleção de modelo que busca equilibrar a complexidade do modelo com a qualidade do ajuste aos dados observados. A fórmula para calcular o AIC é dada por:

$$AIC = 2k - 2\ln(L)$$

onde k é o número de parâmetros estimados no modelo e $\ln(L)$ é o logaritmo natural da máxima verossimilhança do modelo. Um menor valor de AIC indica um modelo mais preferível, sugerindo que ele alcança uma boa adequação aos dados com uma quantidade razoável de parâmetros.

O BIC, similarmente ao AIC, é usado para seleção de modelo, mas inclui uma penalidade mais severa para o número de parâmetros. A fórmula para o BIC é:

$$BIC = \ln(n)k - 2\ln(L)$$

onde n é o número de observações e k é o número de parâmetros no modelo. Assim como o AIC, um menor valor de BIC é preferível, indicando um equilíbrio ótimo entre simplicidade e poder explicativo.

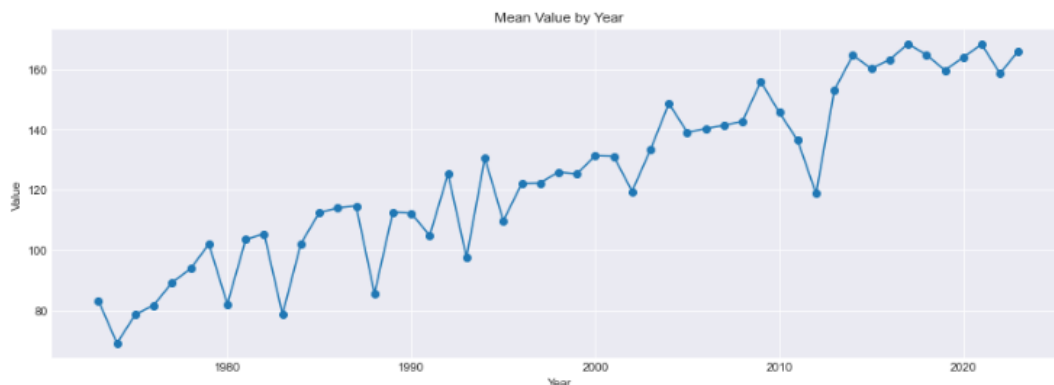
Em modelos multiníveis, utilizaremos a Máxima Verossimilhança Restrita e não mais a Máxima Verossimilhança para estimar o AIC e BIC, mas a estrutura do cálculo permanece igual.

4 Resultados

4.1 Análise Exploratória

Antes de iniciar a modelagem propriamente dita, é essencial definirmos com clareza quais variáveis climáticas serão utilizadas como covariáveis. Além disso, é importante investigar os efeitos temporais e estaduais nas predições. Para tanto, realizaremos uma análise exploratória dos dados para identificar quais variáveis preditoras apresentam correlação significativa com a variável resposta. Neste momento, vamos observar a evolução da produtividade no decorrer dos anos:

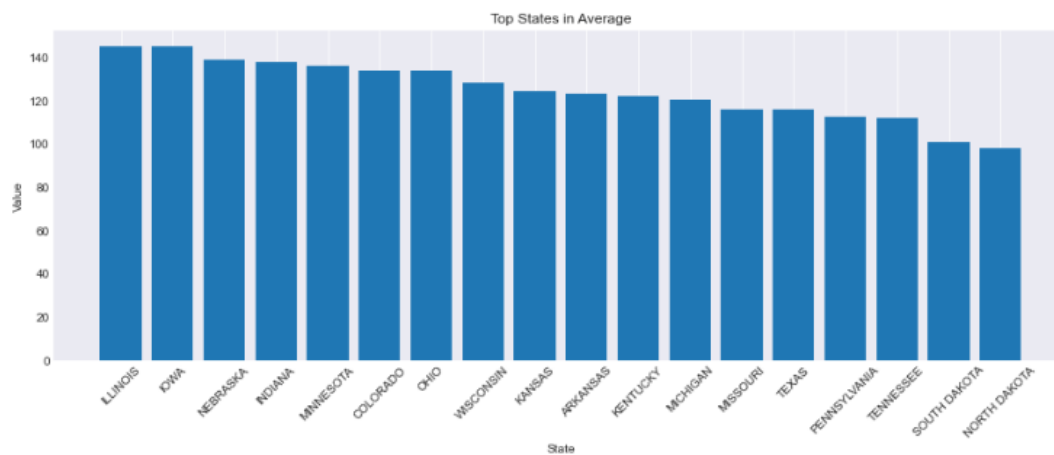
Figure 3: Produtividade por ano



É possível notar uma evolução clara no decorrer dos anos dessa produtividade, possivelmente devido a melhores técnicas de cultivo e evoluções tecnológicas. Isso será interessante para a etapa em que testaremos modelos multiníveis.

Agora, vamos entender o comportamento da produção em cada estado:

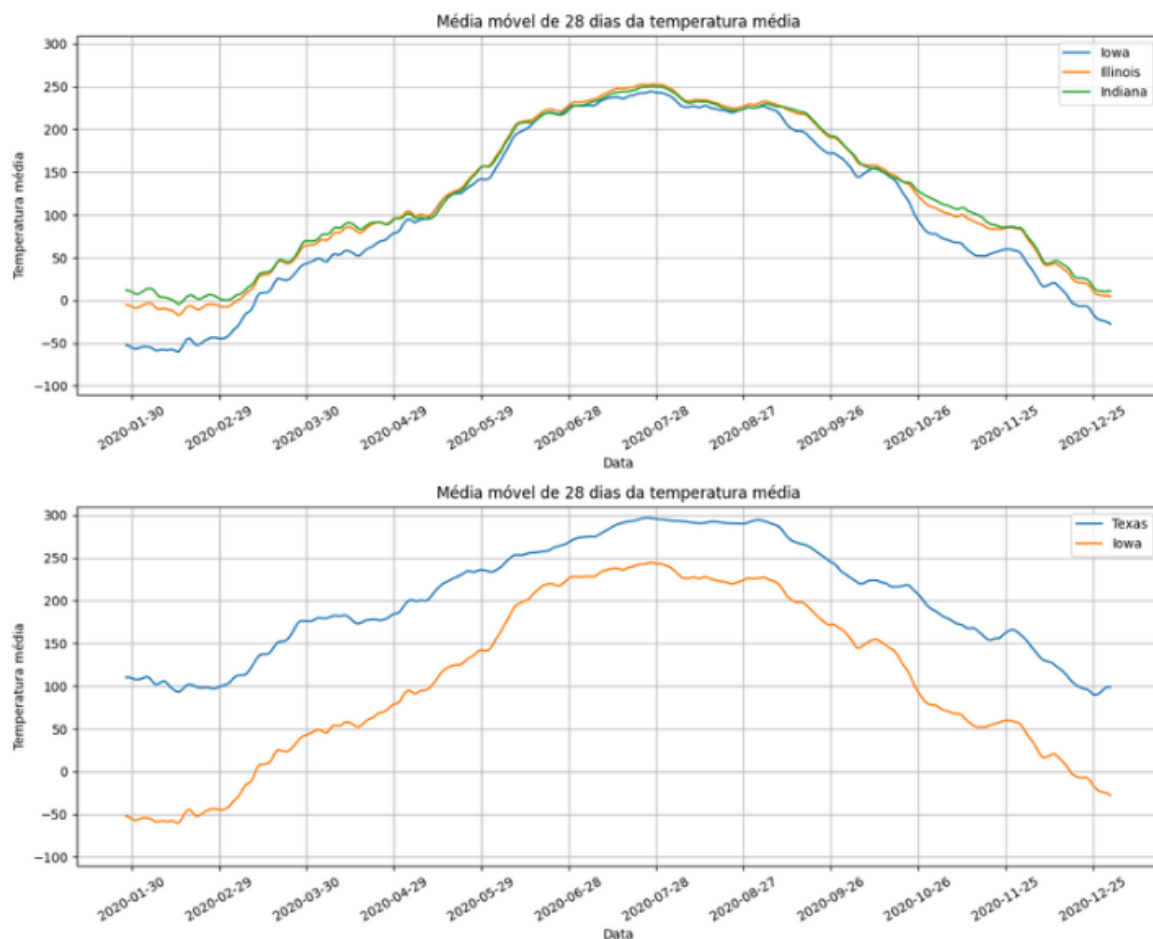
Figure 4: Produtividade média por estado



Há uma diferença considerável entre os estados. Entre os estados concentrados no *Corn Belt* (como Illinois, Iowa e Indiana), região ao centro-norte do país que o clima é considerado melhor para o plantio do milho, a produtividade é mais alta em relação a outras localizações.

Agora, vamos partir para uma análise mais focada nas variações climáticas. Uma hipótese inicial que tínhamos é que o clima em cada estado impactaria em sua produtividade final. Vamos conferir se há uma diferença considerável entre o clima de localizações diferentes:

Figure 5: Média móvel do clima nos estados



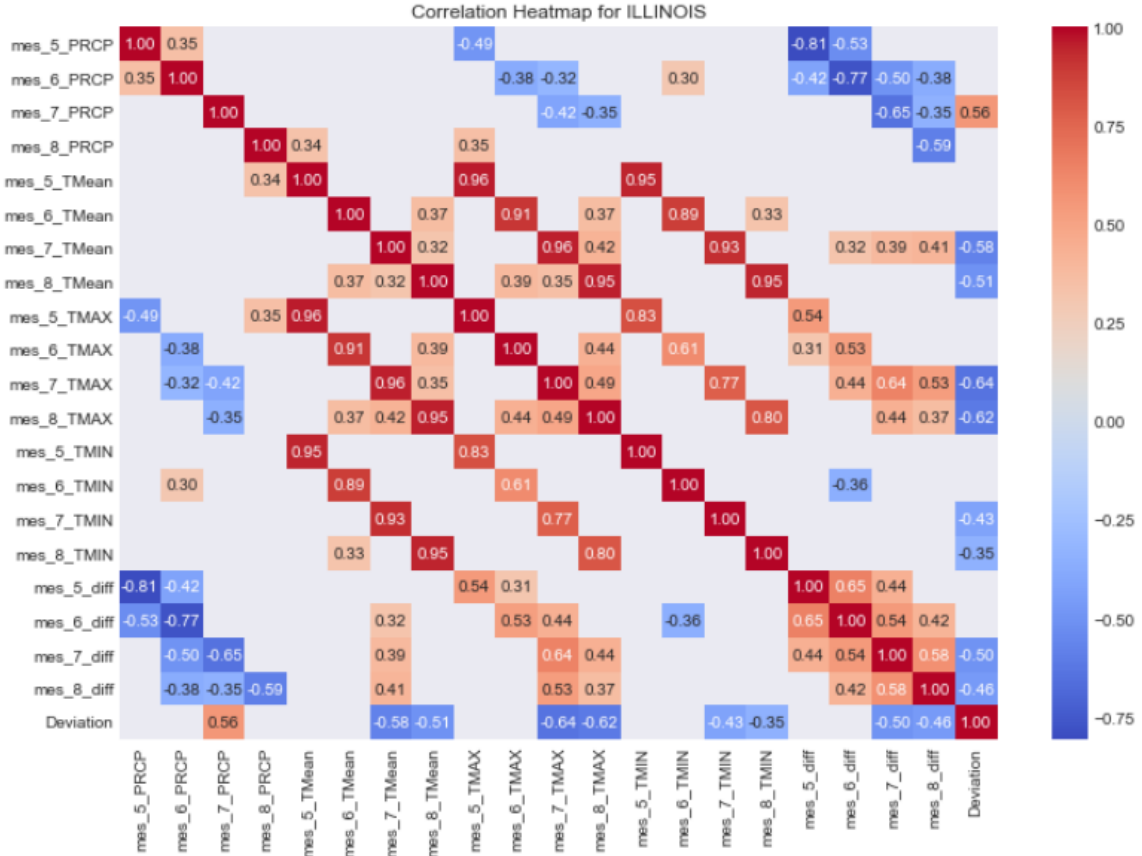
No primeiro gráfico, conferimos a média móvel da temperatura média de 3 estados do *Corn Belt*, e é possível notar que se parecem bastante. No 2º gráfico, comparamos Iowa (localizada no centro-norte do país) com o Texas, estado ao sul dos EUA. Nessa ocasião, notamos que há uma diferença de temperatura considerável, mas ambos possuem uma tendência semelhante (maiores temperaturas no meio do ano e menores no começo/fim).

Agora, precisamos entender exatamente quais covariáveis climáticas utilizar. E, para isso, vamos calcular a correlação de Pearson que cada uma delas possui com a variável resposta, o *yield*. Entretanto, um problema que teria ao realizar essa comparação é o claro efeito anual que há na produtividade, tornando um pouco mais difícil de ver como cada variável afeta a produtividade. Por isso, especificamente para essa análise, iremos substituir nossa variável resposta *yield* pelo desvio da tendência do *yield*, assim vamos conseguir "anular" o efeito temporal em nosso *heatmap*. Para calcular esse desvio, realizo o seguinte cálculo:

$$dev_yield = \frac{yield}{trend} - 1$$

Onde dev_yield é nosso desvio, $yield$ é nossa produtividade projetada e $trend$ é a tendência anual. A seguir, o *heatmap* do estado de Illinois, mas o dos outros 17 estados podem ser encontrados [aqui](#). Foi plotado apenas correlações com valor absoluto maior que 0.3, para facilitar a visualização.

Figure 6: Correlação de Pearson para Illinois



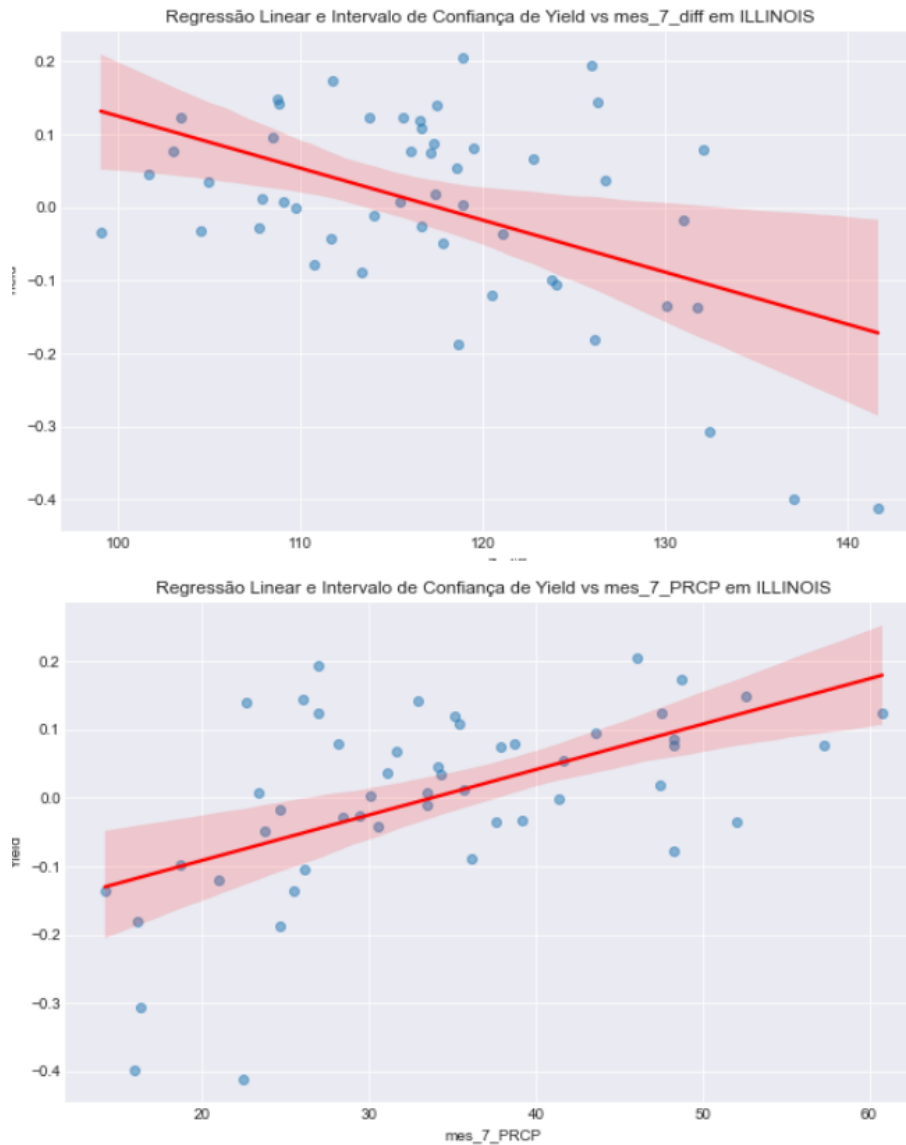
Essa visualização nos dá diversas informações que podem ser úteis. O mês 5, independente da variável, não mostra uma correlação tão interessante para a variável resposta, não importando o estado analisado. Isso confirma informações que foram vistas na revisão de literatura, onde vimos que o mês de maio não é crucial para a produtividade.

Além disso, as variações de temperatura parecem ter uma boa correlação com o desvio, como nesse caso $TMAX$ 7 possui coeficiente -0.64 . $diff$ 7 também tem uma boa correlação, mas pode haver caso de multicolinearidade na hipótese de usarmos essas duas variáveis, que possuem entre si coeficiente 0.64 .

A precipitação tem uma correlação interessante também no mês 7, informação que também vimos na revisão de literatura, que pontuou junho e julho como importantes para a produtividade.

Como $PRCP$ 7 e $diff$ 7 são variáveis que se destacaram, vamos ver como a nuvem de pontos se comporta quando relacionamos cada uma dessas variáveis com o desvio:

Figure 7: Nuvem de pontos - *diff*, *PRCP*



No caso da diferença de temperaturas, parece que valores mais altos prejudicam o *yield*, mas diferenças medianas não nos dá tanta informação. No caso da precipitação, parece haver uma tendência de melhores *yields* quando a precipitação tem um valor um pouco maior, de pelo menos 30. Quando é muito baixa, há quebras de safra (valores reais muito abaixo da tendência).

Com essas informações, conseguiremos modelar os dados com mais precisão. Temos noção de quais variáveis e meses parecem mais importantes, e também temos um direcionamento de como é o comportamento temporal e estadual.

4.2 Modelos ajustados

Nas primeiras modelagens, testarei quais modelos se adequarão melhor aos nossos dados. Para isso, utilizaremos 4 variáveis que se destacaram tanto na literatura quanto na análise exploratória: precipitação nos meses 6 e 7, diferença de temperatura nos meses 7 e 8. Depois disso, faremos testes com outras covariáveis até chegar em um melhor modelo.

4.2.1 Regressão linear múltipla simples

Em um primeiro momento, vamos trabalhar com uma metodologia simples: modelar uma regressão linear. A principal característica deste modelo é que ele assume que as observações são independentes entre si e que têm a mesma variância (homoscedasticidade). Chamaremos esse modelo de **Modelo 1**. Será mais simples de entender, porém não captaremos a possível correlação entre medições feitas no mesmo ano ou variações na influência de variáveis ao longo do tempo. Os seguintes resultados foram reportados:

Table 1: Resultados do Modelo 1

	Variável	Estimação	Erro padrão	AIC	BIC	MSE
Modelo 1	Intercepto	-3242.19	100.65	8222	8255	447.28
	ano	1.6871	0.049			
	mes_6_PRCP	0.1746	0.059			
	mes_7_PRCP	0.3521	0.072			
	mes_7_diff	0.1833	0.084			
	mes_8_diff	-0.3408	0.070			

Uma modelagem sem a variável *ano* também foi testada, porém, como previsto pelas informações na análise exploratória, os resultados foram inferiores e muito menos satisfatórios - obtemos AIC 8978, BIC 9006 e erro quadrático 1021.4. Por isso, desconsideramos e focamos em comparar modelos mais condizentes com nossas informações conquistadas.

No **Modelo 1**, cada unidade adicionada às variáveis preditoras altera a produtividade conforme indicado na coluna "Estimação". O intercepto inicial de -3242.19, embora possa parecer atípico, representa a produtividade estimada em 1973, com todas as outras variáveis nos seus valores base. Este valor baixo é compensado pelo ajuste anual proporcionado pela variável *ano*, que adiciona cerca de 1.6871 à produtividade a cada ano subsequente, capturando assim as tendências de longo prazo que as variáveis climáticas não abrangem. Um aumento na precipitação nos meses 6 e 7 está associado a um leve aumento na produtividade, enquanto a variação de temperatura nos meses 7 e 8 apresenta um efeito oposto, o que sugere ser um efeito não captado tão bem pelo modelo.

Em relação ao erro padrão das variáveis, obtemos um valor razoavelmente maior para o intercepto, o que pode indicar uma incerteza significativa na previsão da média geral. A variável "ano" apresenta um erro bem mais baixo, e temos uma estimativa mais precisa. Por fim, as variáveis climáticas possuem um erro moderado. Vamos conseguir comparar melhor nossa incerteza na próxima sessão, na presença de outros modelos.

4.2.2 Modelos multiníveis com intercepto variáveis

Nosso **Modelo 2** será um modelo multinível com o intercepto variável por ano. Isso permite que o intercepto capture variações na variável *yield* que podem ser específicas de cada ano, independentemente do estado. Além disso, o modelo irá tratar tanto o estado como as variáveis climáticas como efeitos fixos. Isso significa que o modelo vai estimar um único coeficiente para cada estado e cada variável climática, assumindo que o impacto desses fatores é constante através dos diferentes anos. Os resultados são reportados a seguir:

Table 2: Resultados do Modelo 2

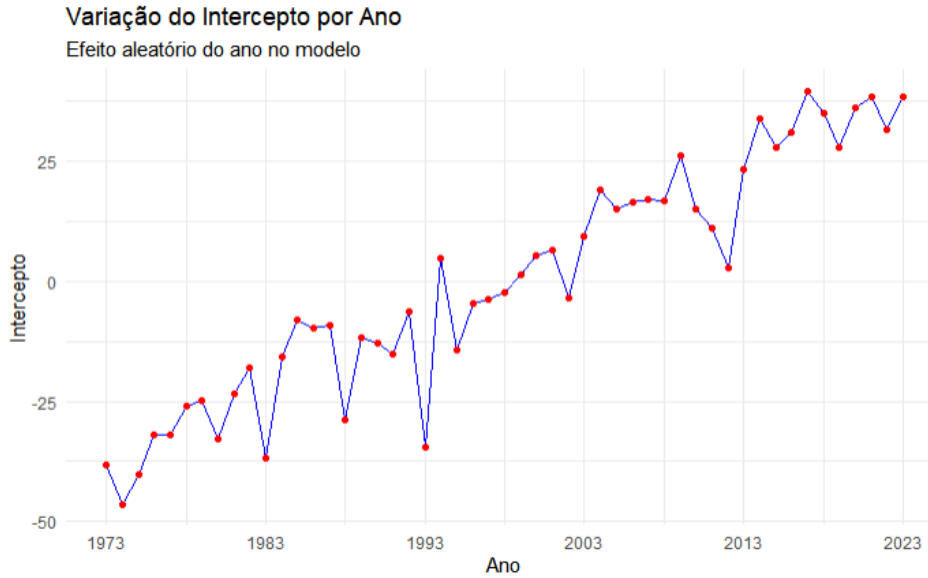
	Variável	Estimação	Erro padrão	AIC	BIC	MSE
Modelo 2	Intercepto	178.125	15.486	7756	7872	208.21
	Estado[COLORADO]	39.131	5.255			
	Estado[ILLINOIS]	19.033	2.999			
	Estado[INDIANA]	10.936	3.004			
	Estado[IOWA]	19.406	3.010			
	Estado[KANSAS]	10.250	3.541			
	Estado[KENTUCKY]	-3.416	3.035			
	Estado[MICHIGAN]	-0.540	3.036			
	Estado[MINNESOTA]	14.155	3.036			
	Estado[MISSOURI]	-7.606	2.994			
	Estado[NEBRASKA]	28.338	3.681			
	Estado[NORTH DAKOTA]	-10.783	3.663			
	Estado[OHIO]	8.267	2.997			
	Estado[PENNSYLVANIA]	-11.496	3.070			
	Estado[SOUTH DAKOTA]	6.965	3.857			
	Estado[TENNESSEE]	-14.633	3.023			
	Estado[TEXAS]	-1.519	3.040			
	Estado[WISCONSIN]	4.165	3.027			
	mes_7_diff	-0.273	0.096			
	mes_8_diff	-0.248	0.069			
	mes_6_PRCP	0.111	0.051			
	mes_7_PRCP	0.166	0.068			

Obtemos um ótimo avanço em relação ao **Modelo 1**, pois nosso AIC diminuiu quase 500 pontos e BIC 400 pontos, enquanto o MSE foi para 208.21. Isso retrata como cada ano possui fatores específicos que afetam a produtividade final, hipótese que também tínhamos ao analisar os dados. Os coeficientes de cada estado nos concede informações interessantes: estados com maior produtividade média possuem coeficientes mais altos, enquanto estados com baixa produtividade possuem coeficientes mais baixos. Isso acontece porque a interpretação dessa estimativa leva em conta o Arkansas (primeiro estado em ordem alfabética) como estado de referência, ou seja, dizemos que Illinois tem 19.033 unidades maior do que no Arkansas, e a mesma ideia segue para as outras estimativas. É curioso o caso do estado de Colorado, que não está situado no *Corn Belt* e possui uma produtividade média razoável, mas foi estimado com um coeficiente bastante alto.

Em relação as variáveis climáticas, conseguimos ter um efeito que corresponde mais a realidade. O aumento na precipitação aumenta o *yield*, enquanto o aumento da diferença de temperaturas abaixa o valor da variável resposta. O mês 7 tem coeficiente maior em ambos os casos, o que retrata a importância desse mês.

Em relação ao erro dos parâmetros fixos estimados, temos que o erro padrão das variáveis climáticas é muito parecido com o que vimos no **Modelo 1**, porém um pouco menos em 3 dessas 4 variáveis, mas agora ajustado para as variações anuais. Sobre o erro do intercepto de efeito aleatório, obtemos valor 24.84, sugerindo um impacto considerável entre os grupos. Como estamos estimando apenas um efeito aleatório, em tese essa incerteza será menor do que em modelos futuros, que usaremos coeficientes variáveis. Vamos ver como se comporta essa variação anualmente:

Figure 8: Variação Do Intercepto por Ano



É interessante notar como essa visualização parece o gráfico na Figura 2, sugerindo que a variação significativa nos interceptos ano a ano indica que há fatores temporais ou anuais específicos que influenciam fortemente a produtividade do milho.

Nesse momento, vamos abrir espaço para outra abordagem: o **Modelo 3** será outro modelo multinível, mas com o intercepto variável por estado. Essa abordagem reconhece que diferentes estados podem ter condições de base distintas para a produção de milho, como variações em solo, práticas agrícolas locais, ou microclimas. É uma suposição interessante, pois, como visto, estados situados no *Corn Belt* possuem maior produtividade. Abaixo, os resultados:

Table 3: Resultados do Modelo 3

	Variável	Estimação	Erro padrão	AIC	BIC	MSE
Modelo 3	Intercepto	-2967.146	91.209	7851	7889	267.22
	ano	1.581	0.042			
	mes_6_PRCP	0.142	0.048			
	mes_7_PRCP	0.186	0.067			
	mes_7_diff	-0.254	0.086			
	mes_8_diff	-0.335	0.063			

É possível notar que os resultados são melhores do que o **Modelo 1**, mas não superam o **Modelo 2** quando comparamos o AIC, BIC e MSE. Com essa abordagem, podemos concluir que o efeito estadual existe e é considerável, porém a questão temporal se mostrou mais impactante ainda.

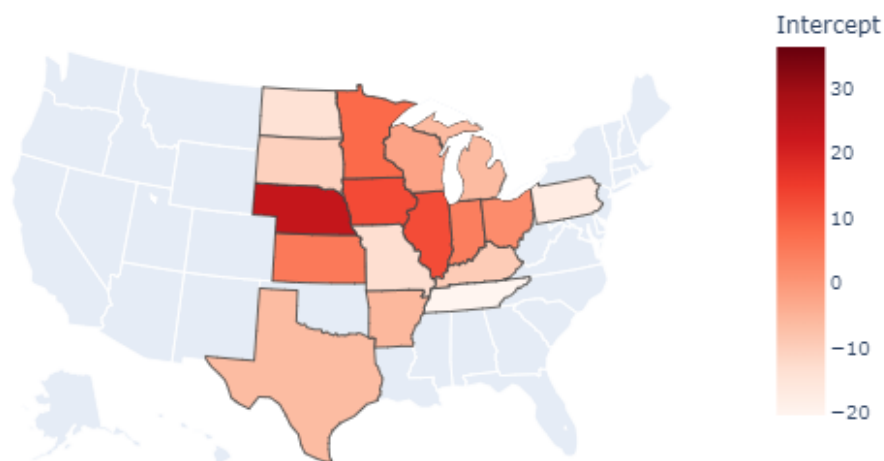
Apesar disso, o **Modelo 3** registrou erros padrão menores para as variáveis climáticas, o que sugere que esse modelo captura melhor essa variabilidade específica entre os estados. Entretanto, essa diferença é ligeiramente pequena, e o **Modelo 2** proporciona um melhor equilíbrio entre ajuste aos dados e penalização pela complexidade do modelo.

O erro padrão do efeito aleatório foi 14.97, relativamente menor do que o efeito aleatório quando a variável ano é o intercepto. Isso indica que, embora haja variabilidade

entre os estados, essa variabilidade é relativamente menos incerta ou mais consistente entre as observações em comparação com a variabilidade anual.

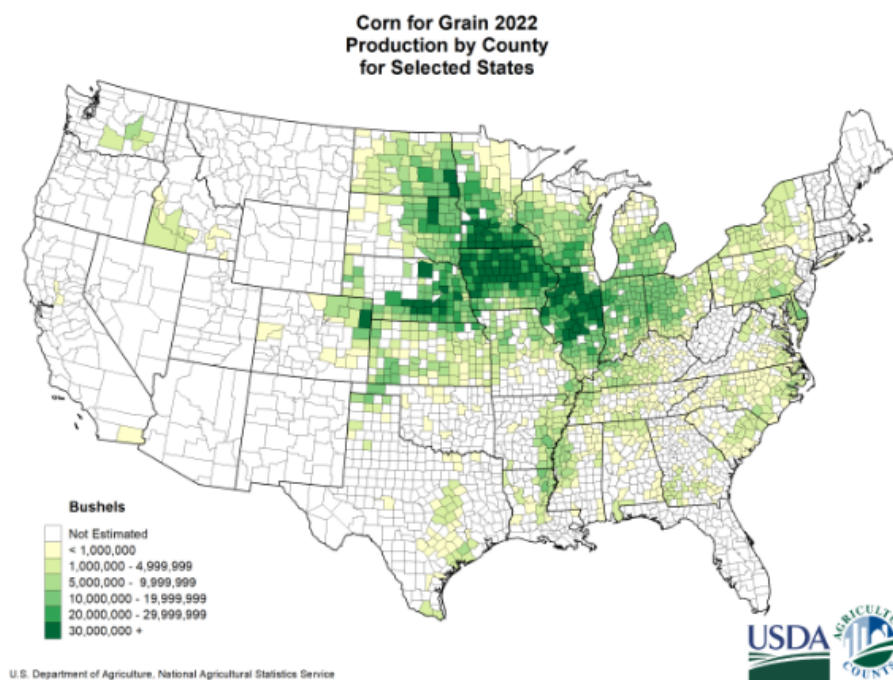
Com auxílio de um mapa, vamos plotar as variações nesse intercepto por estado:

Figure 9: Variação Do Intercepto por Estado



Olhando essa visualização, podemos notar que interceptos são maiores nos estados do *Corn Belt*, que concentram a maior produção de milho no país. O Colorado, curiosamente possui o maior intercepto entre os estados, mesmo não estando naquela região central. Quando comparamos com dados de produção total disponibilizados pelo USDA, notamos a semelhança entre os gráficos:

Figure 10: Produção por County



Com esses gráficos, concluímos que estados que produzem mais também possuem maior produtividade. Isso é interessante porque mostra que possivelmente, por terem uma importância maior na produção do cereal, é necessário que a produtividade nesses estados seja a maior possível.

Na próxima subseção, vamos testar modelar com coeficientes variáveis, ao passo em que mantemos o intercepto (ano) também variável.

4.2.3 Modelos Multiníveis com Intercepto e Coeficientes Variáveis

Na última subseção, concluímos que a variação no intercepto por ano foi essencial para obtermos bons resultados. Em uma análise mais complexa, podemos considerar que os efeitos climáticos em cada ano possuem efeitos diferentes, ou seja, uma alteração em tempos atuais pode ter significado diferente do que uma alteração em épocas distantes. Como vimos na análise exploratória que o mês de julho era crucial para a produtividade, vamos conferir como as variáveis nesse mês se comportam com efeito aleatório. Nesse sentido, em nosso **Modelo 4** escolhemos modelar as variáveis climáticas de julho com efeitos aleatórios para capturar essas variações anuais. Mantemos o Estado, a precipitação de junho (*PRCP_6*) e a diferença de temperatura de agosto (*Diff_8*) como efeitos fixos. Assim, nosso modelo se adapta para explorar mais detalhadamente como os fatores climáticos de julho interagem com cada ano específico, buscando uma compreensão ainda mais refinada de suas influências na produtividade.

Table 4: Resultados do Modelo 4

	Variável	Estimação	Erro padrão	AIC	BIC	MSE
Modelo 4	Intercepto	196.09	15.322	7672	7811	172.68
	Estado[COLORADO]	46.59	4.970			
	Estado[ILLINOIS]	19.13	2.830			
	Estado[INDIANA]	11.22	2.838			
	Estado[IOWA]	21.13	2.855			
	Estado[KANSAS]	14.00	3.350			
	Estado[KENTUCKY]	-3.57	2.860			
	Estado[MICHIGAN]	0.92	2.874			
	Estado[MINNESOTA]	16.44	2.885			
	Estado[MISSOURI]	-6.72	2.830			
	Estado[NEBRASKA]	32.81	3.487			
	Estado[NORTH DAKOTA]	-6.12	3.479			
	Estado[OHIO]	8.34	2.835			
	Estado[PENNSYLVANIA]	-10.74	2.909			
	Estado[SOUTH DAKOTA]	-2.32	3.652			
	Estado[TENNESSEE]	-14.76	2.853			
	Estado[TEXAS]	0.34	2.870			
	Estado[WISCONSIN]	5.80	2.874			
	mes_6_PRCP	0.096	0.050			
	mes_7_PRCP	0.153	0.073			
	mes_7_diff	-0.42	0.095			
	mes_8_diff	-0.25	0.060			

Conseguimos estimativas ainda mais precisas do que nosso **Modelo 2**, onde o AIC abaixou para 7672 e o BIC abaixou para 7811. Apesar de termos aumentado a complexidade do nosso modelo, que é um ponto a se tomar cuidado, nossas estimativas para cada estado se tornaram mais precisas, com erros padrão menores também. Da mesma forma que o **Modelo 2**, estados com maior produtividade tem valor estimado maior do que os estados com menor produtividade ao compararmos com o Arkansas, estado de referência.

Quando olhamos a incerteza sob nossos coeficientes aleatórios, temos a seguinte tabela:

Table 5: Incerteza sob oas variáveis de efeito aleatório do modelo 4

Variável	Variância	Erro padrão
Intercepto	1283	35.81
mes_7_PRCP	0.0394	0.198
mes_7_diff	0.2999	0.1732

A incerteza neste modelo é maior porque há mais parâmetros aleatórios sendo estimados (intercepto e coeficientes). Isso naturalmente resulta em uma maior variabilidade nos parâmetros estimados do nosso modelo. Apesar disso, os resultados foram muito interessantes.

4.2.4 Outras abordagens e resultados

Quando testamos essa mesma abordagem com outras variáveis climáticas, obtemos variações muito insignificantes no AIC, MSE e BIC. Isso acontece porque, como são variáveis climáticas, estão fortemente correlacionadas (principalmente TMIN, TMAX, Tmean e Diff). Ao adicionarmos mais variáveis no modelo, como a precipitação no 8º mês, também temos diferenças mínimas no AIC e MSE, enquanto o BIC começa a ter valores mais elevados.

Outras abordagens também foram testadas, como elevar variáveis ao quadrado para podermos acentuar valores mais altos, porém não tivemos melhora significativa. Apesar de valores mais altos de temperaturas possuírem maior impacto na produção, temos uma baixa amostragem desses dados que passam de um limite fora do razoável, o que impacta em nosso ajuste.

Em suma, o nosso 4º e último modelo foi o melhor encontrado. Essa abordagem mais complexa foi preferível para modelarmos a produtividade do milho, pois os efeitos climáticos sob a ótica temporal são diferentes. É algo perfeitamente razoável com a realidade, onde podemos imaginar que melhores práticas agrícolas ao longo do tempo fizeram com que surgissem melhores adaptações a efeitos climáticos. Abaixo, é reportado os resultados gerais de todos os modelos:

Table 6: Resultados gerais dos modelos

Modelo	R^2	AIC	BIC	MSE
Modelo 1	0.625	8222	8255	447.25
Modelo 2	0.789	7756	7872	208.21
Modelo 3	0.794	7851	7889	267.22
Modelo 4	0.766	7672	7811	172.68

Considerando que estamos avaliando a produtividade com base apenas nos dados climáticos, obtivemos resultados interessantes. O **modelo 4** apresentou um bom desempenho em termos de MSE, AIC e BIC, além de mostrar tendências regionais e temporais muito interessantes. Apesar de aumentarmos a complexidade dos nossos modelos, as métricas indicaram melhorias, permitindo captar nuances na produtividade que poderiam passar despercebidas em modelagens mais simplistas. No entanto, um ponto a ser destacado é que o R^2 deste modelo não foi o mais alto, porém, ainda é muito próximo dos outros modelos multiníveis, e razoavelmente maior que o modelo mais simples. Ele reporta uma boa capacidade explicativa, chegando a quase 80%.

5 Conclusão

O trabalho, de modo geral, nos concedeu informações valiosas quanto ao plantio do milho nos EUA. Concluímos que o fator temporal se mostrou extremamente relevante para estimar essa produtividade. Observamos que avanços nas técnicas de agricultura têm impactado positivamente essa produtividade ao longo do tempo. Além disso, conseguimos compreender o efeito específico de cada variável climática nos diversos meses do ciclo de plantio, assim como a correlação de cada uma dessas variáveis com a variável de interesse. Embora modelos mais simples possam ser úteis, eles muitas vezes não capturam as nuances específicas que a nossa questão de pesquisa exige. Por outro lado, os modelos multiníveis utilizados neste trabalho demonstraram ser extremamente eficazes, permitindo uma análise mais detalhada e precisa.

Uma limitação no trabalho é a redundância das informações climáticas e a consequente multicolinearidade, que impacta negativamente as previsões. Para abordar essa questão, é promissor considerar a inclusão de dados não relacionados ao clima em trabalhos futuros. O USDA, que disponibilizou a variável resposta, possui diversas informações detalhadas sobre o plantio anual, e a inclusão desses dados podem enriquecer nossas análises, levando a estimativas mais acuradas.

Outro aspecto interessante é que o início do plantio varia entre os estados, sendo um pouco mais tardio nos estados ao norte e mais cedo nos estados ao sul. Embora essa diferença seja mínima, existe a possibilidade de que os meses críticos considerados não sejam exatamente os mesmos para cada região. Capturar essa variação pode melhorar significativamente o ajuste dos dados, utilizando meses diferentes como covariáveis para cada estado.

Além dessas melhorias, é interessante considerar, para trabalhos futuros, uma melhor ponderação do clima utilizando os sensores mais importantes de cada estado. Atualmente, selecionamos a média geral das informações climáticas do estado, o que inclui sensores em áreas que não são de plantio. Com uma melhor definição dessas áreas, conseguiríamos selecionar apenas sensores mais importantes.

References

- Andrew Gelman and Jennifer Hill. *Data Analysis Using Regression and Multi-level/Hierarchical Models*. Cambridge University Press, 2006.
- Andrew Gelman, Jennifer Hill, and Aki Vehtari. *Regression and Other Stories*. Cambridge University Press, 2020.
- Vijaya R. Joshi, Maciej J. Kazula, Jeffrey A. Coulter, Seth L. Naeve, and Axel Garcia y Garcia. In-season weather data provide reliable yield estimates of maize and soybean in the us central corn belt. *International Journal of Biometeorology*, 65:489–502, 2021. doi: 10.1007/s00484-020-02039-z.
- National Oceanic and Atmospheric Administration. NOAA.gov. <https://www.noaa.gov/>, 2024. Accessed: 2024-05-28.
- R. H. Shaw and J. E. Newman. Weather stress in the corn crop. *Cooperative Extension Service, Michigan State University*, November 1985. MSU is an affirmative-action equal-opportunity institution.
- United States Department of Agriculture. USDA.gov. <https://www.usda.gov/>, 2024. Accessed: 2024-05-28.