The impact of external factors on honey bees productivity in the USA

Advance Multivariate Statistics Project

Gallo Luigi - Longo Angelica - Rizzi Melissa

December 2024

1 Introduction

Every day, we are reminded of the crucial role that bees play in sustaining life on Earth. Bees are vital not only for their honey production but, more importantly, for their role as pollinators. Through pollination, they contribute significantly to biodiversity and the stability of ecosystems, supporting the growth of countless plant species, including many that are essential for human agriculture and food supply. However, bee populations face increasing threats from a variety of environmental and anthropogenic factors, such as climate change, pesticide exposure, habitat loss, and diseases. These threats not only jeopardize the survival of bees but also have profound implications for agricultural productivity and ecological balance.

For this study a rich dataset is used that covers honey production per colony in different states in the United States, along with atmospheric and pollution variables. Using advanced multivariate statistical techniques, we aim to explore and quantify the influence of these variables on bee productivity. Understanding the relationship between external variables, such as temperature, precipitation, and pesticide use, and bee productivity is critical to gain valuable information on the health and productivity of colonies under various environmental conditions and to develop strategies to mitigate these threats and ensure the sustainability of bee populations.

2 Data and Preliminary Explorations

The dataset used was derived from official sources that provide comprehensive records of honey production[1] (e.g. yield per colony, total production, price per pound) by

[1] `https://www.nass.usda.gov/Surveys/Guide_to_NASS_Surveys/Bee_and_Honey/index.php`

state and year, as well as environmental and pollution data, including air quality[2], atmospheric conditions[3], natural disasters[4] and pesticide usage[5].

2.1 Data Transformation and New Variables

The raw data set required several cleaning and transformation steps to prepare it for analysis. The following are details of each preprocessing step:

- The state was standardized to ensure consistent formatting (e.g. typos or inconsistencies in state names).
- Variables recorded in inconsistent units, such as area in square miles or temperature in Fahrenheit, were converted, respectively, to square kilometers and Celsius.
- A new variable, colperkm2, was calculated by dividing the number of colonies by state area (area_km^2).
- A new variable, nAllNeonic_km2, was calculated as the sum of all pesticides divided by state area (area_km^2).
- Data from 1994 to 2016 were combined to create a comprehensive dataset grouped by state and year. This involved ensuring consistent variable names, handling missing values, and merging across multiple files.

Data is not available for all states, particularly for certain years. Smaller states such as New Hampshire and Rhode Island are missing, likely due to the limited presence of beekeeping activities in those regions. Similarly, Alaska lacks data, which may be attributed to its harsh climate and minimal apicultural practices. This absence of information highlights regional disparities in beekeeping prevalence and the challenges in collecting consistent data across all areas.

2.2 Initial Exploratory Data Analysis

The following graph (Figure 1) shows the total number of colonies in the United States over time, from 1991 to 2021. A sharp decline can be observed in the early years, followed by a more volatile trend in subsequent years.

[2] https://aqs.epa.gov/aqsweb/airdata/download_files.html#Annual

[3] https://www.ncei.noaa.gov/access/monitoring/climate-at-a-glance/statewide/time-series/1/tmax/12/0/1996-2024?base_prd=true&begbaseyear=1901&endbaseyear=2000

[4] https://www.fema.gov/disaster/declarations

[5] https://water.usgs.gov/nawqa/pnsp/usage/maps/county-level/

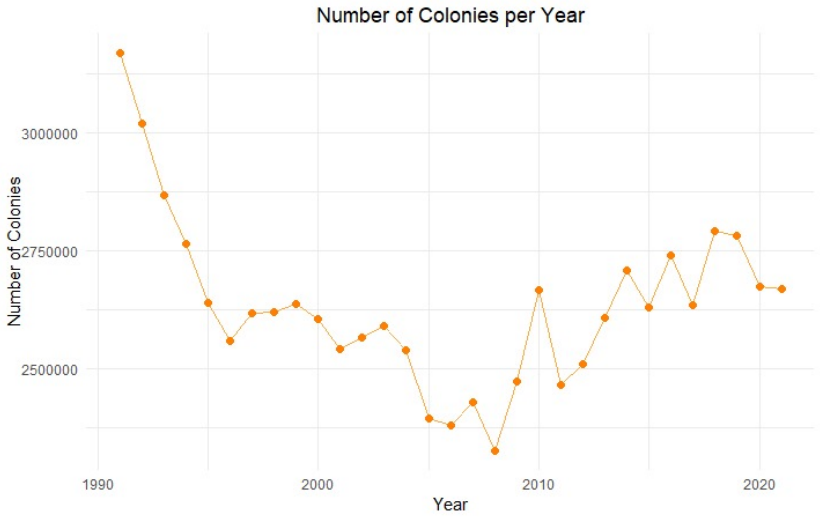

Figure 1: Total Number of colonies in USA Over Time

There is a distinct trend observable for each state, both in the number of honeybee colonies (Figure 2) and in the average honey production per colony (Figure 3). These trends highlight the varying dynamics of beekeeping and honey production across regions. In some states, the yield per colony has shown a consistent decline over the years, potentially reflecting challenges such as environmental stressors, disease prevalence, or shifts in agricultural practices. Conversely, other states have demonstrated more stable patterns or even slight increases.

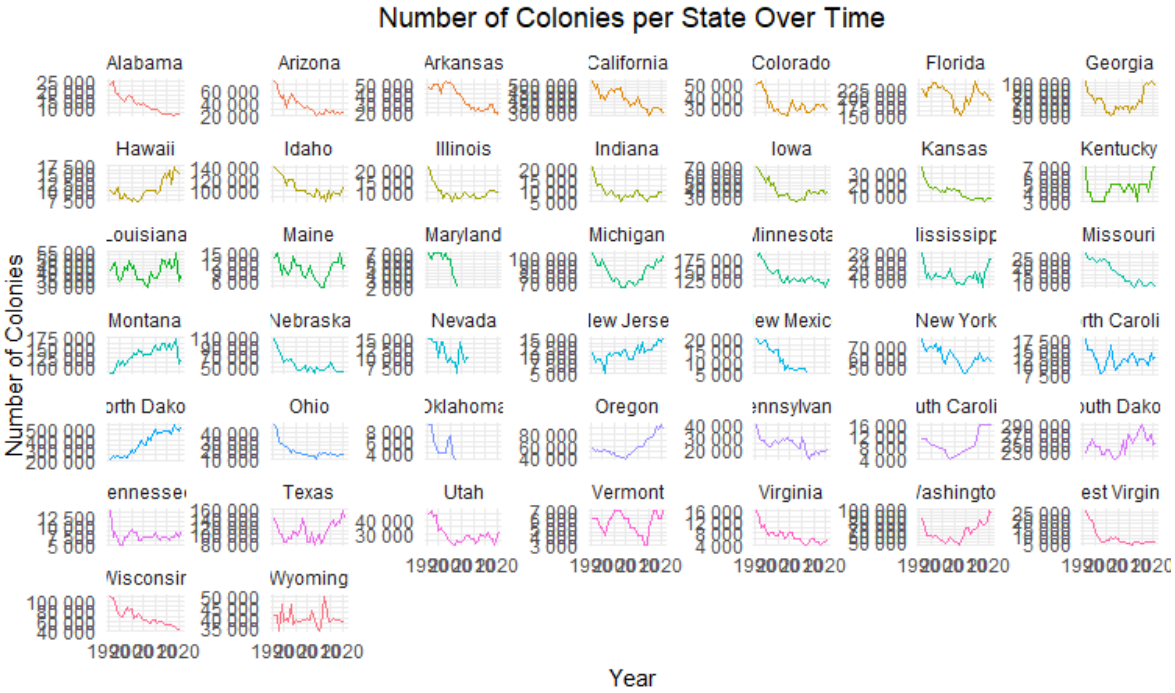

Figure 2: Number of colonies per State Over Time

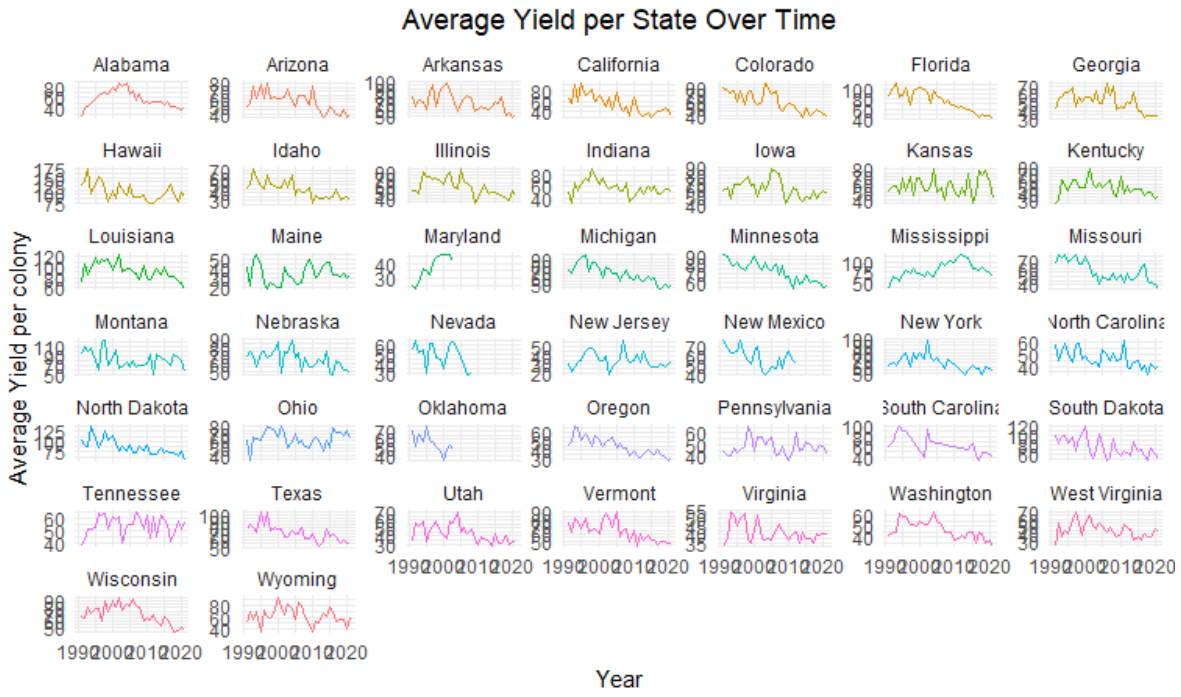

Figure 3: Average Yield per State Over Time

Overall, the data underscore the complexity of regional differences in both colony numbers and productivity.

There is also a noticeable variation in honey production across states over time. In general, states with higher production levels, such as California (see Figure 4), have experienced a decline, while production in other states has tended to stabilize around the average. This trend highlights a convergence toward more uniform production levels across states.

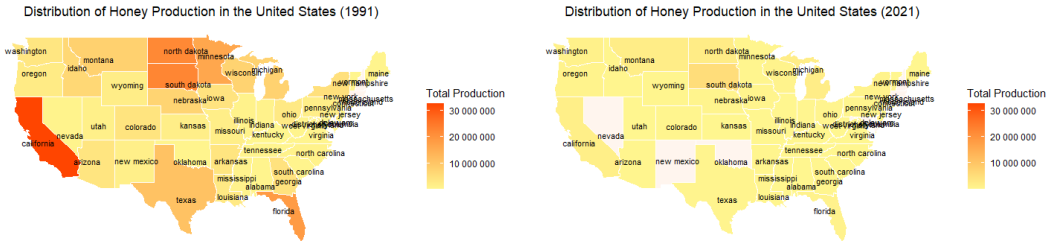

Figure 4: Total production 1991 - 2021

Finally, the price trend can be observed, with evidence of inflation and exponential growth over the years considered. However, the focus of this analysis is on the impact of environmental conditions on honey production, rather than on economic variables. Therefore, price will not be a key variable in our analysis.

Figure 5: Price per lb Over Time

3 Clustering

After gaining an initial understanding of the data through exploratory analysis, the next step was to investigate whether states could be grouped into meaningful categories based on their honey production and other related variables. The selected ones are:

- The number of colonies per square kilometer.
- The price per pound of honey.
- The yield per colony.

The aim is to identify patterns and similarities between states, which may reveal underlying trends in beekeeping practices and economic factors.

3.1 K-Means (2021)

The analysis focuses on the most recent year for which data is available, namely 2021. Data is available for 40 out of the 50 states, while the remaining 10 states—Alaska, Rhode Island, New Hampshire, Connecticut, Delaware, Massachusetts, Oklahoma, Nevada, New Mexico, and Maryland—are missing.

Figure 6 provides a visual representation of the distribution of states across the three selected variables, offering insights into regional patterns and trends.

The method chosen for the analysis is the k-means clustering algorithm, which is a widely used unsupervised learning technique for partitioning data into a predefined number of clusters. The algorithm assigns each data point to the nearest cluster center and iteratively adjusts the cluster centers to minimize the variance within each cluster.

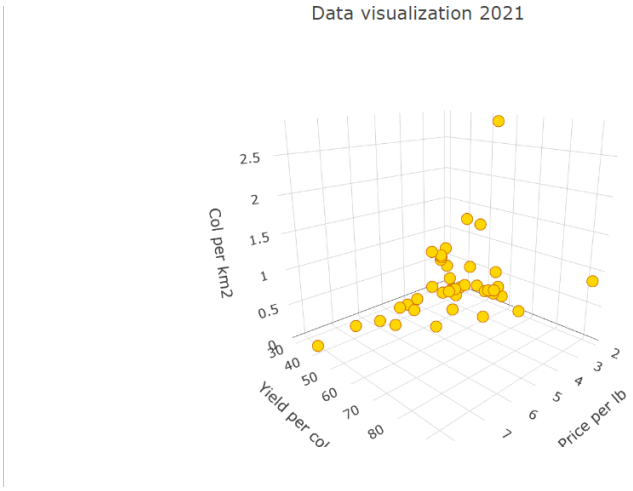

Figure 6: Data Visualization 2021

To determine the optimal number of clusters, two methods will be used:

- **Elbow Method**: involves plotting the sum of squared distances from each point to its assigned cluster center (within-cluster sum of squares) for a range of cluster numbers. The optimal number of clusters is typically chosen at the "elbow," where the rate of decrease in the sum of squares slows down.

  Figure 7 does not exhibit a clear and distinct "elbow," which typically indicates the optimal number of clusters. However, upon closer inspection, the plot seems to suggest that the most reasonable choice for the number of clusters is 4. Further analysis using other methods, such as the silhouette method, can help confirm this choice.

- **Silhouette Method**: assess the quality and cohesion of the clusters. The silhouette score measures how similar each data point is to its own cluster compared to the nearest neighboring one. It ranges from -1 to 1, where a higher average silhouette width indicated more distinct and coherent clusters.

  The improvement in silhouette score becomes less pronounced beyond k = 4, suggesting diminishing returns when increasing the number of clusters and confirming 4 as the optimal amount.

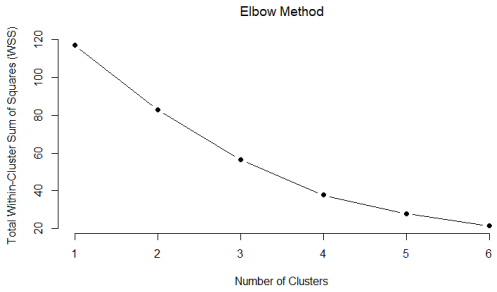

Figure 7: Elbow Method

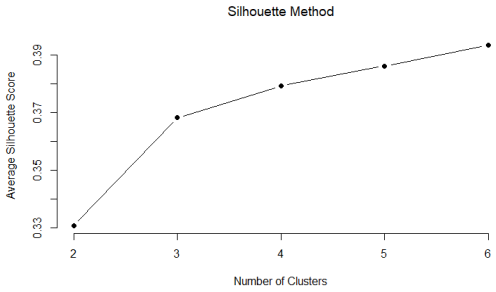

Figure 8: Silhouette Method

The size of the dataset, consisting of 40 statistical units and 3 variables, supports this decision. A higher number of clusters would risk overfitting or creating overly fragmented groupings, while fewer clusters would fail to capture the inherent structure in the data. A 3D Plot of the clusters is shown in Figure 9.

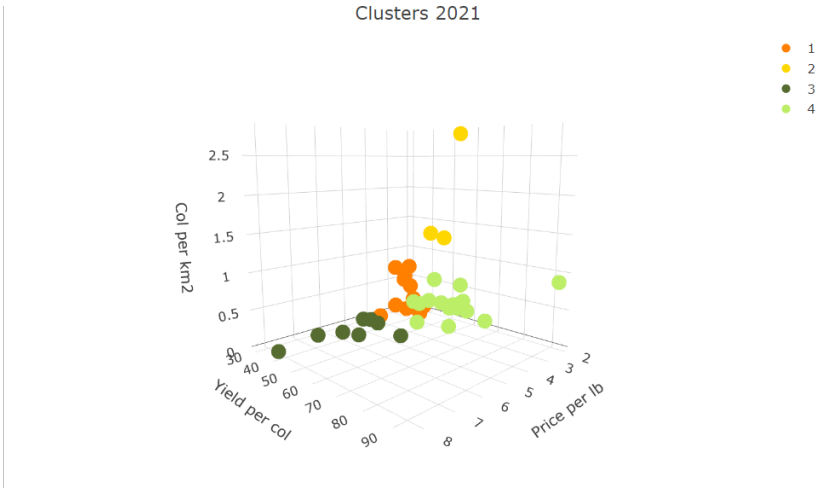

Figure 9: Cluster 3D (2021)

The silhouette plot (Figure 10) provides a more focused visualization of the clustering quality for the chosen k = 4. The silhouette width, represented on the y-axis, measures how well each data point fits within its assigned cluster compared to the next closest one. The plot is divided into four colored regions, each corresponding to one of the clusters identified by the K-means algorithm.

The clusters are mostly compact and well separated, but cluster 4 shows a slightly wider range of silhouette widths, with several points having lower values and some are even negative, showing some possible misclassification. This may reflect outliers that do not fit well with any cluster and for that reason will be examined further using outlier detection techniques.

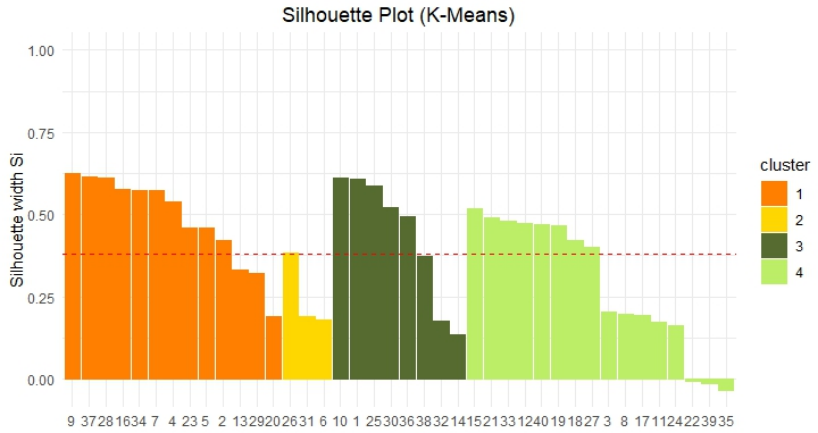

Figure 10: Silhouette Plot

3.2 Outlier Analysis

To ensure the robustness and reliability of the clustering and mixed-effects modeling results, an outlier analysis was performed. Outliers were identified using Mahalanobis Distance and Robust Mahalanobis Distance to capture extreme observations that deviate significantly from the multivariate structure of the dataset.

3.2.1 Mahalanobis Distance Analysis

In Figure 11, the Mahalanobis Distance was calculated to detect multivariate outliers based on the relationships between the key variables (priceperlb, yieldpercol, and colperkm2). The Mahalanobis Distance considers the covariance structure of the data and measures how far each observation lies from the center of the distribution.

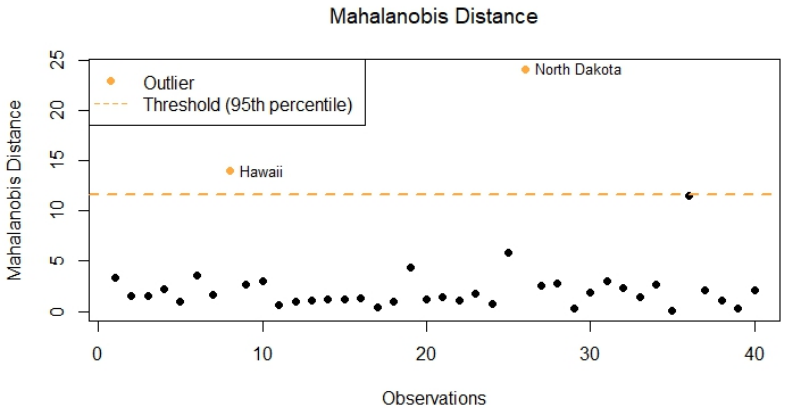

Figure 11: Mahalanobis Distance

North Dakota significantly exceeds the threshold, indicating an extreme deviation from the general pattern of the data. Similarly, Hawaii falls just above the threshold, although to a lesser extent. These two outliers exhibit values for the selected variables that do not align with the overall trend of the dataset. For example, the data show that North Dakota has a number of colonies per square kilometer much higher than the average.

3.2.2 Robust Mahalanobis Distance Analysis

While the traditional Mahalanobis Distance is sensitive to extreme values, the Robust Mahalanobis Distance (Figure 12) mitigates this issue by using robust estimates of the covariance matrix, providing a more reliable measure in the presence of extreme observations.

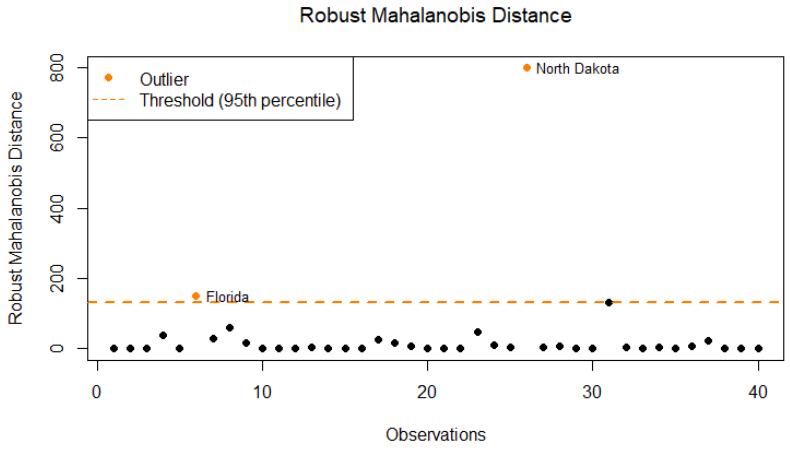

Figure 12: Robust Mahalanobis Distance

The robust approach confirms North Dakota as an extreme outlier, while swapping Hawaii and Florida as outliers just above the threshold.

3.2.3 Distance-Distance Plot

The Distance-Distance Plot (Figure 13) compares the traditional Mahalanobis Distance with the Robust Mahalanobis Distance.

North Dakota checks again its status as an extreme outlier under both methods, while Hawaii and Florida are highlighted as outliers only in one of the two methods (traditional and robust respectively).

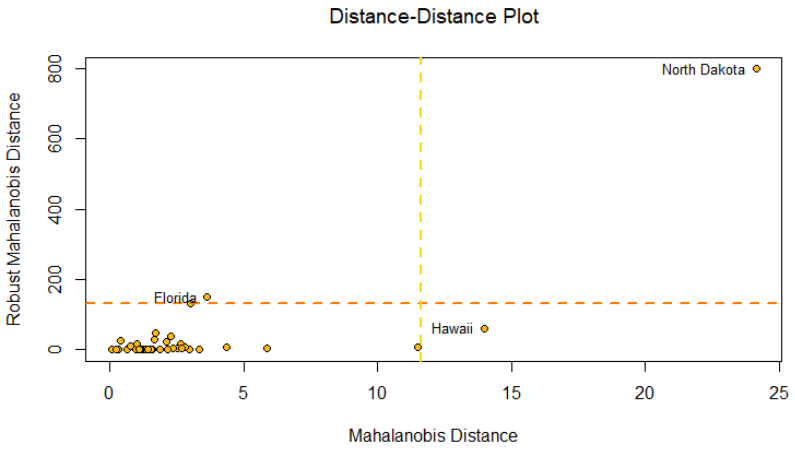

Figure 13: Distance-Distance Plot

3.3 Robust K-Means Clustering

The previous outlier analysis using Mahalanobis distances identified extreme values that could influence traditional clustering methods, motivating the use of a trimmed K-Means clustering approach. Trimmed K-Means is robust to outliers because it excludes a portion of the most extreme observations when computing centroids, ensuring that the clustering results reflect the core data structure more accurately.

3.3.1 Silhouette Analysis for Robust K-Means (2021)

To determine the optimal number of clusters, the Silhouette Method was again applied to evaluate cluster quality for k=3 (Figure 14), k=4 (Figure 15), and k=5 (Figure 16).

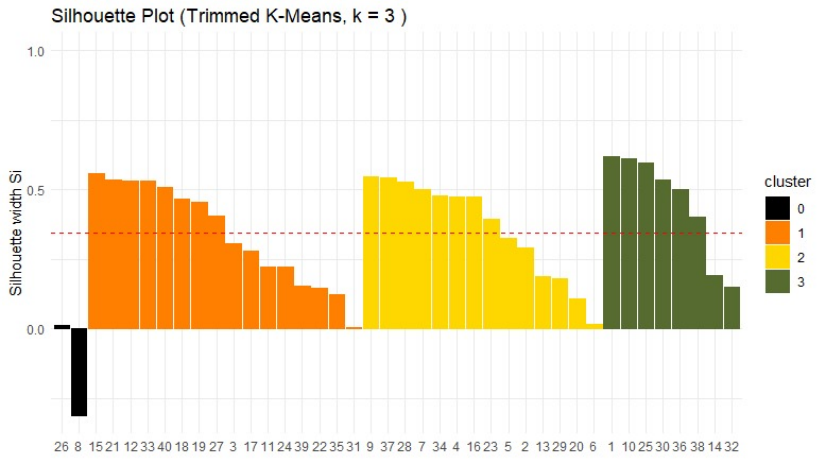

Figure 14: Silhouette Plot k=3

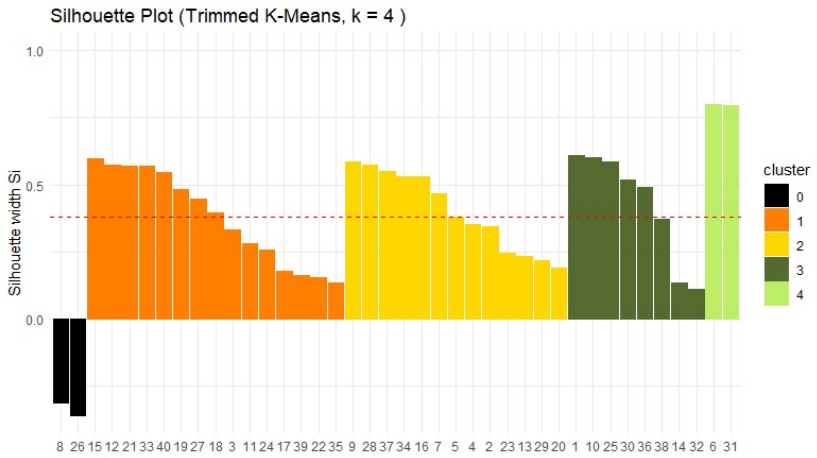

Figure 15: Silhouette plot k=4

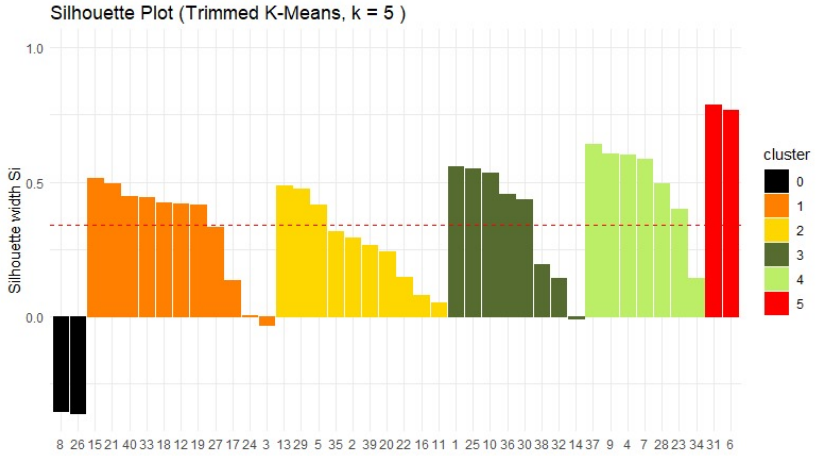

Figure 16: Silhouette plot k=5

After careful consideration, k=4 was confirmed as the best choice. The silhouette widths are predominantly positive with fewer points near the boundary between clusters, the four clusters exhibit good balance and their sizes appear consistent. While k=5 improves the silhouette widths for some observations, it introduces lower silhouette values and even some negative ones. This indicates that splitting into five clusters introduces unnecessary complexity without a significant gain in cluster quality.

Figure 17 illustrates the distribution of the variables in three-dimensional space, where it is evident that the outliers are represented in black and excluded from the identified clusters. This visual representation emphasizes the distinction between the outliers and the majority of the data points, which form distinct clusters based on the selected variables.

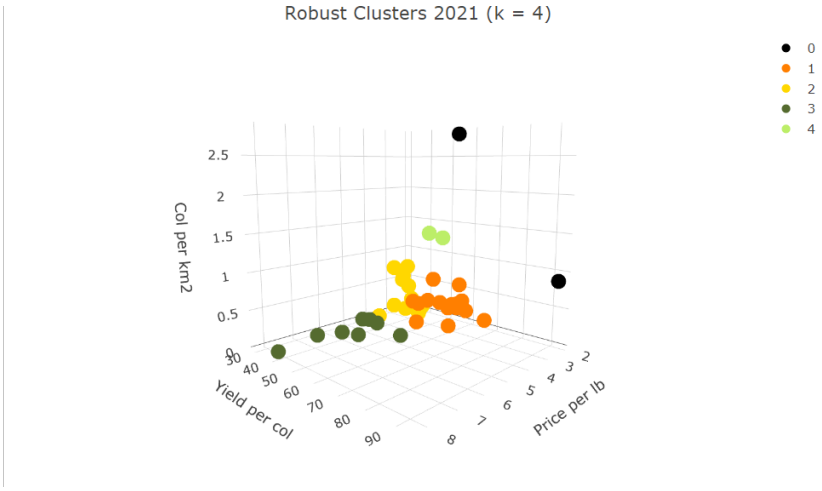

Figure 17: 3D Cluster 2021

3.3.2 Robust K-Means (1991)

The analysis aims to compare the most recent year available (2021) with the earliest year in the dataset (1991), for which clustering was performed using the robust k-means method with $k = 4$. This approach allows for a meaningful comparison between the clusters identified in the two time periods. Below is the 3D plot of the clusters obtained for 1991 (Figure 18).

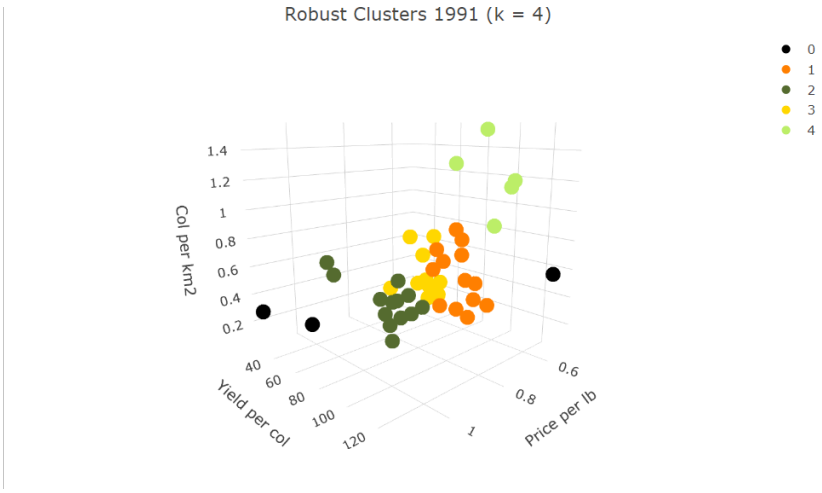

Figure 18: 3D Cluster 1991

The figure 18 shows that, in terms of general characteristics, four similar clusters are formed for both years, although with different sizes and varying centroid values. For instance, the price is significantly higher in 2021 compared to 1991.

3.3.3 Centroid Comparison

The centroid results for k=4 provide a detailed summary of the cluster characteristics in terms of Price per lb, Yield per col, and Col per km^2.

- **2021 Centroids** (Table 1): The table summarizes the cluster characteristics based on the selected variables. Cluster 1 and Cluster 2 show similar prices per pound, but differ in yield per colony and colony density. Cluster 3 is distinct with the highest price per pound and the lowest colony density. Cluster 4 stands out with the highest colony density and the lowest price per pound, but includes only two states, indicating it is an outlier group.

Cluster	Price per lb	Yield per col	Col per km^2	N states
1	2.78	54.9	0.285	15
2	2.77	35.5	0.321	13
3	5.83	42.9	0.0904	8
4	2.34	46.5	1.29	2

Table 1: Cluster's centroids 2021

- **1991 Centroids** (Table 2): Prices per pound in all clusters are significantly lower in 1991, reflecting historical market conditions. Cluster 4 again stands out with the highest colony density and yield per colony, although it includes a slightly larger number of states compared to 2021. Clusters 1, 2, and 3 show closer alignment in colony densities and yields, indicating less pronounced differentiation in 1991 compared to the more distinct patterns observed in 2021. This suggests a growing divergence in production and market dynamics over time.

Cluster	Price per lb	Yield per col	Col per km^2	N States
1	0.575	71.5	0.388	13
2	0.763	48.7	0.211	13
3	0.568	42.4	0.321	10
4	0.538	88.2	1.190	5

Table 2: Cluster's centroids 1991

3.3.4 Geographical distribution

The geographic distribution of the clusters for 2021 and 1991, shown in Figures 19 and Figure 20, provides a visual comparison of how clustering patterns evolved over time. The clustering patterns between years are consistent, with eastern states maintaining high colony densities and central/western states exhibiting low densities but higher honey prices, but states experienced a general decline in honey production (yield per colony), particularly in high-density regions.

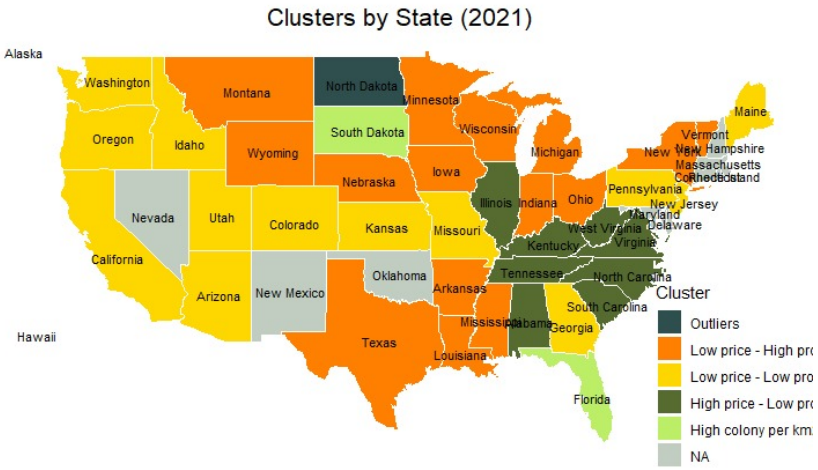

Figure 19: Clusters map 2021

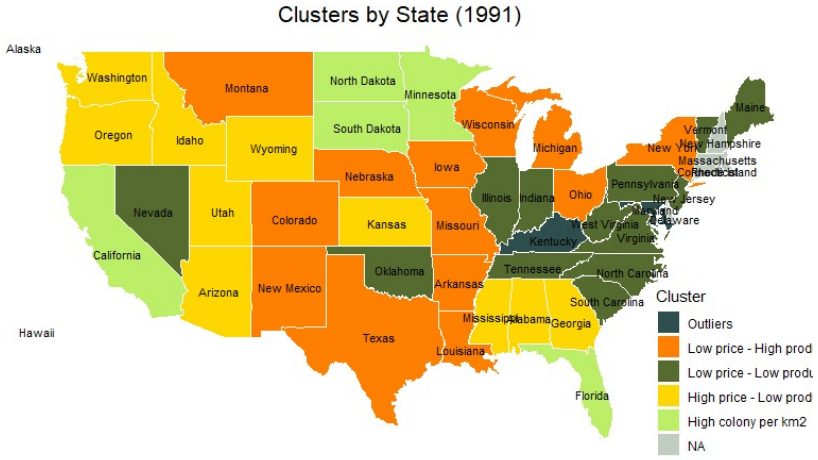

Figure 20: Clusters map 1991

4 Honey Bees' Productivity Analysis

In this chapter the general productivity trends over time and accross states were analyzed, highlighting the need for a model that considers variability between states while assessing fixed effects over time. Figure 21 shows average yield per colony over time for each state, hinting at a generally downward trend.

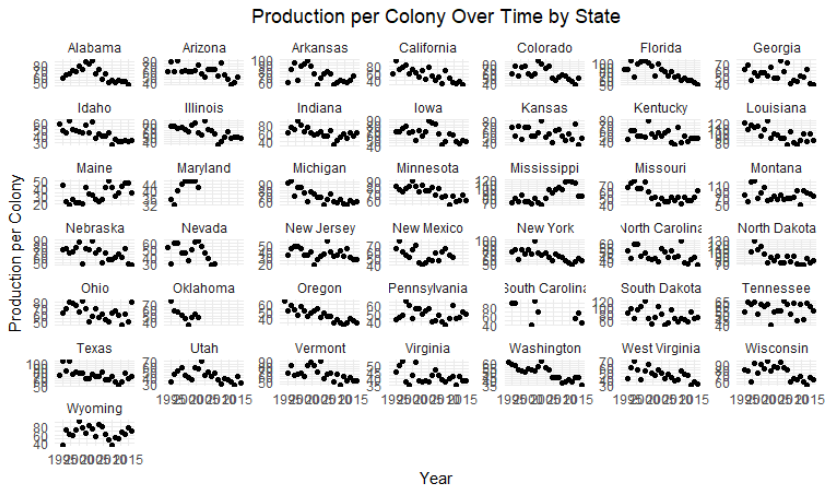

Figure 21: Avg yield over time

4.1 Full Pooling

To further assess this overall pattern, a simple regression model was fitted under the full pooling assumption, where all states follow the same trend with no state-level differences, as depicted in Figure 22

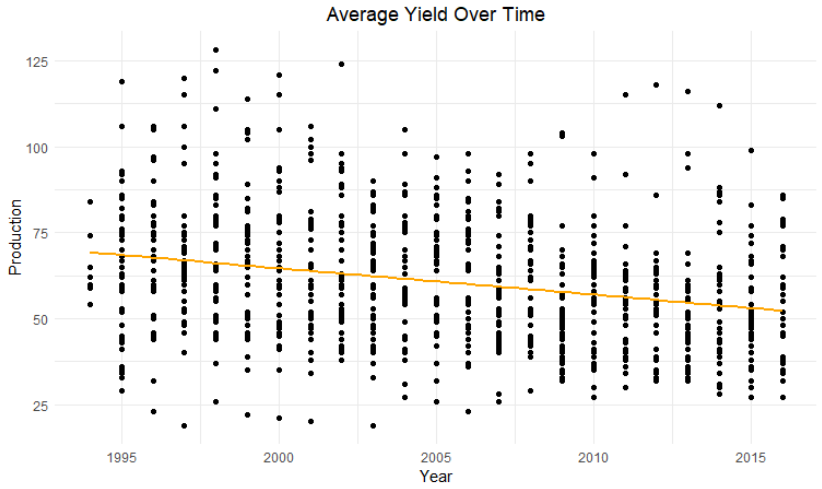

Figure 22: Full pooling

The model confirms a declining trend in productivity over time. However, performance metrics like R^2 (0.06828) and RMSE (18.04565) are relatively poor, indicating that the model is too simplistic and fails to account for important variability among states.

4.2 No Pooling

To explore whether states differ significantly, a no pooling model was applied, fitting separate regressions for each one. Figure 23 shows that some states exhibit steeper declines, while others maintain stable or even slightly increasing trends over time. The R^2 (0.7252) and RMSE (9.8) improve compared to the full pooling model, confirming that state-level differences are non-negligible. The no pooling method is still not the most efficient, it captures inter-state variability but it provides no overarching summary, making it difficult to identify common patterns.

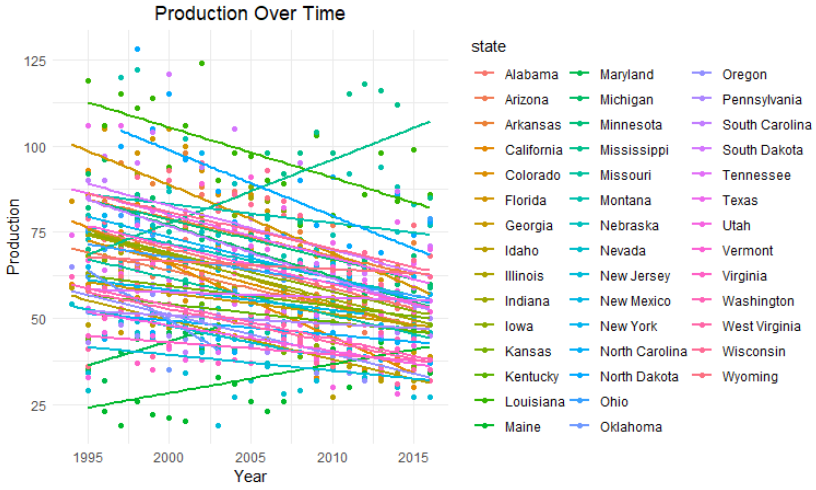

Figure 23: No pooling

4.3 Mixed-effect Model

As a compromise between the overly simplistic full pooling and the overly fragmented no pooling methods, a mixed-effect model is then introduced (Figure 24), including a fixed effect for the global time trend and random intercepts for each state. This final model shows a general downward trend common to all states, with individual that differ in their baseline productivity levels. The fit metrics (such as a reduced RMSE around 10.75) compared to the first model demonstrate that accounting for state-level variability leads to more accurate and interpretable results.

Having established that a mixed-effects model is the most appropriate framework for capturing both the general downward trend in productivity and the variability among states, the next step is to verify that the chosen model meets key statistical assumptions and to confirm that including state-level random effects significantly improves the model.

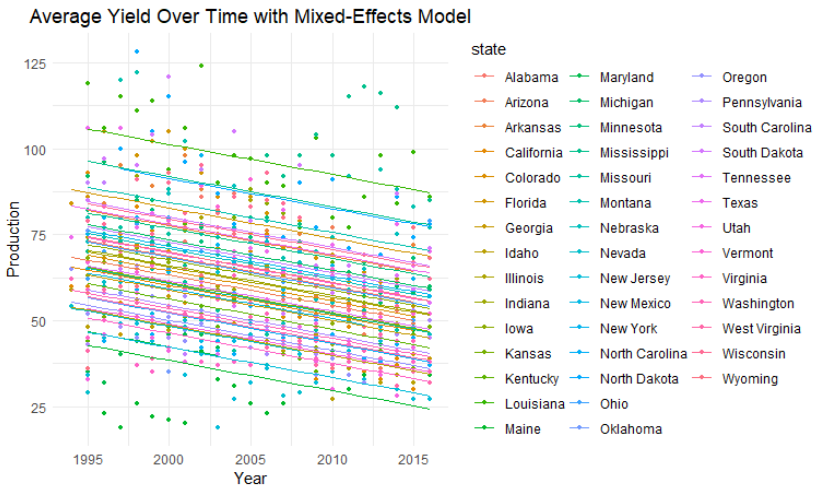

Figure 24: Mixed-effect

4.3.1 Residual Analysis

Figure 25 presents two critical diagnostic plots for the mixed-effects model:

- **Residuals vs. Predicted Values**: The residuals appear randomly scattered around zero, with no evident patterns or trends. This suggests that the assumption of homoscedasticity is reasonably met. There is no visible funneling or curvature, which would indicate problems such as non-linearity or heteroskedasticity.
- **Normal Q-Q Plot**: The residuals follow the theoretical normal line closely, being just a bit sparse at the right end, indicating that the normality assumption for the residuals is not severely violated. Although perfect normality is never guaranteed, the residuals approximate a normal distribution well enough for inference and interpretation.

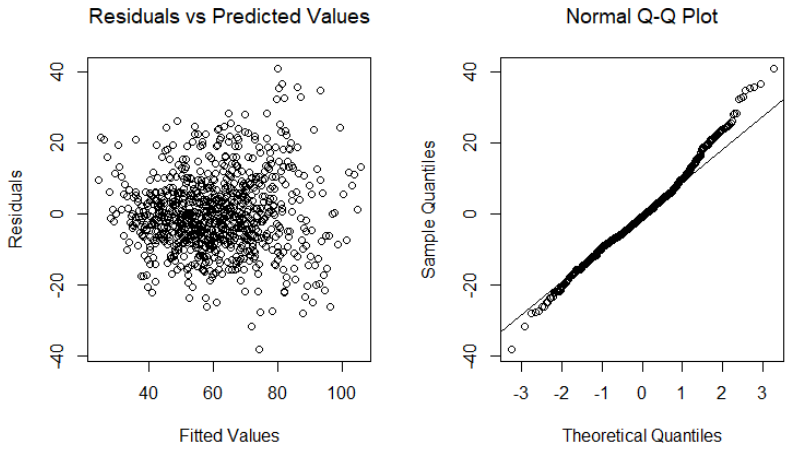

Figure 25: Residuals and Q-Q Plot

4.3.2 Model Comparison Using ANOVA and Information Criteria

Table 3 compares a simple model (without state-level variability) to the full mixed-effects model (with state-level random intercepts) using AIC, BIC, log-likelihood, deviance, and a likelihood ratio test (Chisq and p-value) to confirm the fixed component impact on the model.

Model	**npar**	**AIC**	**BIC**	**logLik**	**Deviance**	**Chisq**	**Pr(>Chisq)**
model_simple	3	7192.9	7207.3	-3593.5	7186.9		
model_full	4	6995.9	7015.1	-3493.9	6987.9	199.04	$< 2.2 \times 10^{-16}$

Table 3: Models comparison

The large and highly significant Chi-square test statistic indicates that the model including random intercepts for states provides a significantly better fit than the simpler model without them. The p-value is effectively zero, making it statistically justified. The full model's AIC and BIC are lower than those of the simple model, reinforcing that the mixed-effects structure better explains the data.

4.3.3 Intraclass Correlation Coefficient (ICC)

The Intraclass Correlation Coefficient provides a measure of how much of the total variation in yield is attributable to differences among states. With an Adjusted ICC of approximately 0.636 and an Unadjusted ICC around 0.582, these values indicate that a sizable portion (over half) of the variation in productivity is due to state-level differences. The ICC supports the idea that states have distinct baseline productivity levels and potentially different responses to environmental conditions.

4.3.4 Visualization of State-Level Effects

Figure 26 (Dotplot of Random Intercepts) further illustrates the importance of considering state-level variation. This plot shows the estimated random intercepts for each state. States with intercepts above zero have baseline productivity levels higher than the global average, while those below zero have lower baseline productivity. The horizontal distribution of points confirms substantial variability in how states deviate from the overall mean and some states stand out as consistently higher or lower in productivity, hinting at unique local conditions.

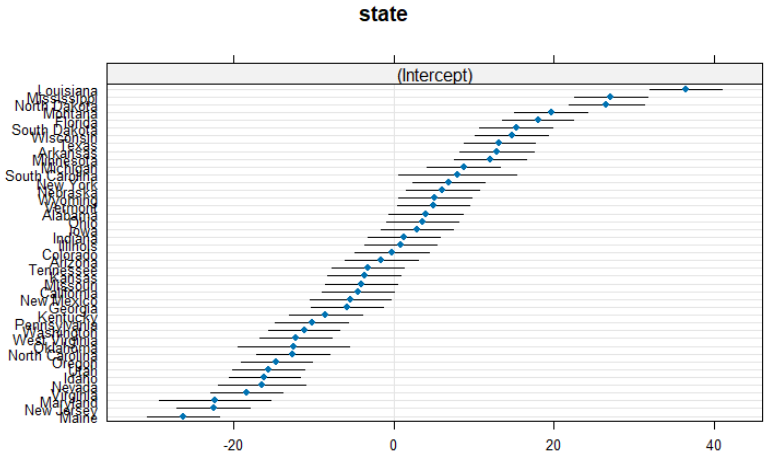

Figure 26: Dotplot

4.3.5 Predicted vs. Observed Values

Figure 27 (Predicted vs. Observed Values) shows how well the model's predictions align with actual data. The points are around the diagonal line, indicating that the model predictions are generally close to the observed values. The spread is also relatively tight around the line, meaning that the mixed-effects approach has enhanced predictive accuracy.

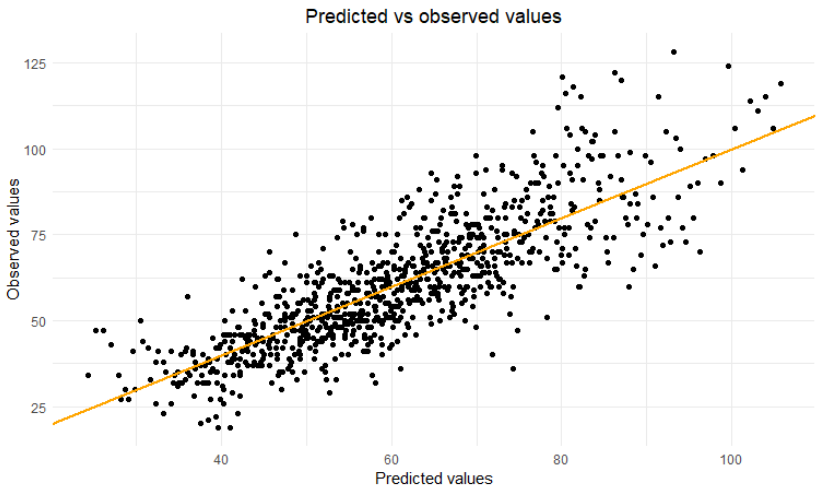

Figure 27: Predicted vs actual values

5 Variables Selection and Outlier Investigation

This chapter focuses on selecting and refining environmental variables that may explain why bee productivity is declining, ranging from pesticide use and air quality measures to climatic and atmospheric conditions, and on identifying and addressing outliers that could distort the analysis. Initially, a wide range of candidate predictors was considered, including:

- **Pesticides (per km²)**: nCLOTHIANIDIN, nTHIACLOPRID, nACETAMIPRID, nTHIAMETHOXAM, nIMIDACLOPRID
- **Air Quality**: Max_AQI, Days_CO, Days_NO2, Days_Ozone, Percent_Good_Days, Percent_Unhealthy_Days, Days_PM2.5, Days_PM10
- **Natural Disasters**: Flood, Fire and Storm_Group
- **Atmospheric Conditions**: Precipitation, Anomaly_prec e Palmer_index
- **Temperature Variables**: Max_Temperature, Min_Temperature

The goal was to identify variables with meaningful relationships to yield without inflating model complexity. To ensure all variables were numeric and consistently scaled, preprocessing steps were taken—such as unit conversions, removing unnecessary columns, and standardizing values.

5.1 Correlation and Multicollinearity Checks

A correlation heatmap (Figure 28) was produced to visualize pairwise relationships.

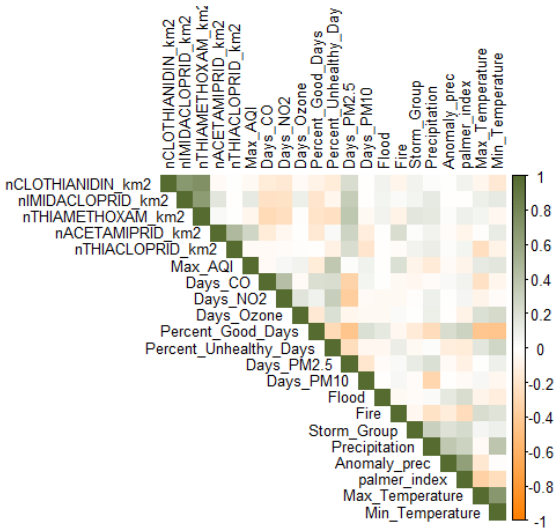

Figure 28: Heatmap

Variable	**VIF**	**sqrt(VIF)**	**sqrt(VIF)> 1.5**
nCLOTHIANIDIN_km2	3.079397	1.754818	TRUE
nIMIDACLOPRID_km2	2.741326	1.654305	TRUE
nTHIAMETHOXAM_km2	3.007544	1.733457	TRUE
nACETAMIPRID_km2	1.686192	1.298535	FALSE
nTHIACLOPRID_km2	1.501066	1.225243	FALSE
Max_AQI	1.452957	1.205393	FALSE
Days_CO	1.471717	1.213988	FALSE
Days_NO2	1.545015	1.242979	FALSE
Days_Ozone	1.170517	1.082842	FALSE
Percent_Good_Days	2.393685	1.546341	TRUE
Percent_Unhealthy_Days	1.799909	1.341619	FALSE
Days_PM2.5	2.100525	1.449318	FALSE
Days_PM10	1.239336	1.112385	FALSE
Flood	1.089191	1.043172	FALSE
Fire	1.257610	1.121383	FALSE
Storm_Group	1.237511	1.112434	FALSE
Precipitation	2.784472	1.668907	TRUE
Anomaly_prec	1.946332	1.394394	FALSE
Palmer_index	2.241917	1.497305	FALSE
Max_Temperature	3.388738	1.840974	TRUE
Min_Temperature	3.960972	1.989993	TRUE

Table 4: VIF values

While no excessively strong correlations were evident, we proceeded with a Variance Inflation Factor (VIF) analysis (Table 4) to ensure that subtle multicollinearity would not undermine our models.

Variables with a $\sqrt{VIF}$ greater than 1.5 were flagged, helping us maintain the ones that contribute unique information without redundancy. A few pesticide variables and certain climate indicators showed higher VIFs or conceptual redundancy. Decisions were made to reduce the set of pesticides to an aggregated measure (nAllNeonic_km2) and to select one temperature measure (e.g. Max_Temperature) to avoid overlap with Min_Temperature. Similarly, Percent_Good_Days was excluded in favor of Percent_Unhealthy_Days for interpretability and due to its overlap with other air quality indicators.

A final check was performed by repeating the same correlation (Figure 29) and VIF analysis, which resulted in a noticeable improvement, as there is no longer any multicollinearity present.

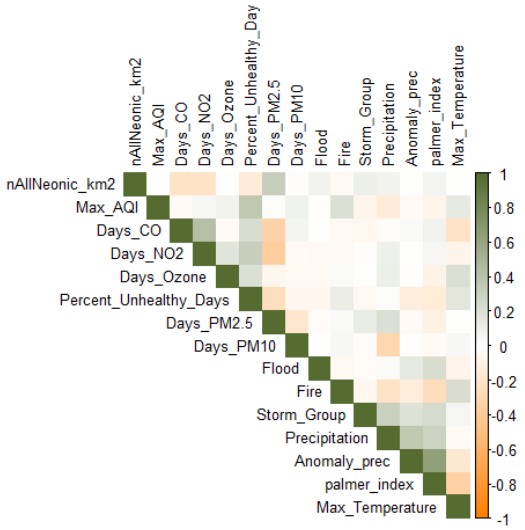

Figure 29: Heatmap after

5.2 Identifying and Addressing Outliers

The following descriptive statistics (Table 5) revealed significant differences between the mean and variance values and their robust counterparts. This could suggest the presence of outliers. This hypothesis appears to be further confirmed by the boxplots (Figure 30), which show that certain variables had extreme values exceeding the interquartile range (IQR).

It is therefore essential to assess whether these extreme values are genuinely anomalous by examining their relationship with the other variables in the dataset. To achieve

Variabile	Mean	Trimmed Mean	Median	Variance	MAD
yieldpercol	60.7095	60.1397	59.0000	349.9021	12.0000
year	2005.1251	2005.0704	2005.0000	40.5257	5.0000
nAllNeonic_km2	0.1490	0.1039	0.0375	0.0914	0.0365
Max_AQI	401.3609	199.9888	196.0000	2075354.5330	28.0000
Days_CO	67.3575	59.5631	10.0000	10528.3239	10.0000
Days_NO2	99.6950	93.9732	62.0000	10044.5478	56.0000
Days_Ozone	323.2637	328.1821	354.0000	4256.6530	10.0000
Percent_Unhealthy_Days	0.6067	0.4620	0.1660	1.0892	0.1660
Days_PM2.5	232.3799	235.9899	276.0000	14998.8399	86.0000
Days_PM10	151.5385	151.4212	89.0000	17995.4524	85.0000
Flood	0.1609	0.1285	0.0000	0.2135	0.0000
Fire	1.0715	0.5162	0.0000	15.0888	0.0000
Storm_Group	0.9441	0.8994	1.0000	1.2340	1.0000
Precipitation	36.4738	36.3344	38.5600	221.2429	10.8800
Anomaly_prec	0.0352	-0.0190	-0.1400	29.8727	3.1600
palmer_index	0.1457	0.1525	0.1000	4.8664	1.4220
Max_Temperature	30.4347	30.4207	30.3300	8.9577	2.3400

Table 5: Descriptive Analysis results

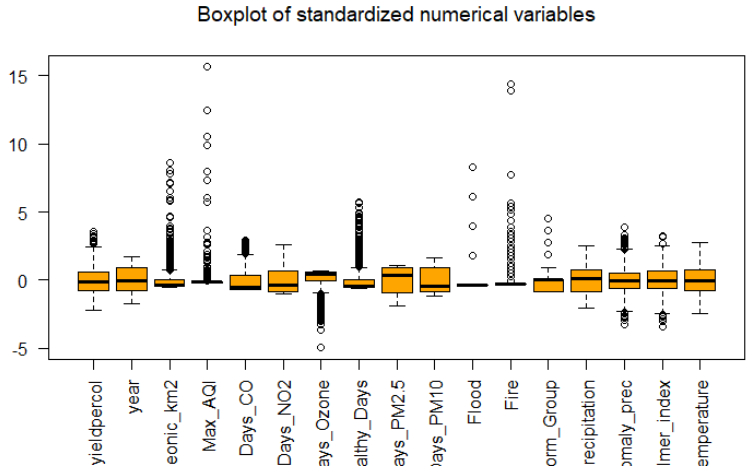

Figure 30: Standardized Boxplot

this, the analysis proceeds by calculating the Mahalanobis distance (Figure 31), which reveals that some values significantly exceed the threshold, corresponding to the 95th percentile, by large margins. This indicates that these values may indeed be outliers and warrant further investigation. Applying a robust version of the Mahalanobis distance (Figure 32) the presence of influential outliers is confirmed, with even more extreme metrics, exceeding 300,000 in some cases.

To prevent these extreme values from skewing the results, the 18 top observations (those above the 98th percentile) were removed. In an attempt to understand why these observations were so far from the rest of the distribution, it was observed that all of them were related to California. This is likely due to the fact that the state experiences pollution levels significantly higher than the average and has a high incidence of fires.

A new round of Malhanobis distances was performed. The picture changed dramatically, with maximum values of robust Mahalanobis distance dropping from over 300,000

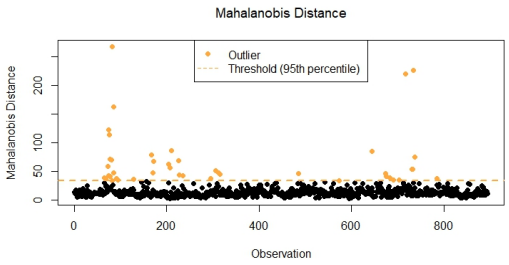

Figure 31: Mahalanobis Distance

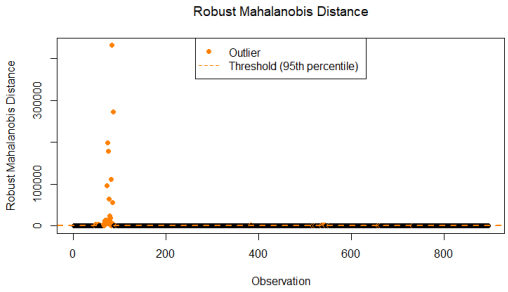

Figure 32: Robust Distance

to around 4,000 for certain variables (Figure 34). These transformations indicate that the data, once stripped of anomalous points, better represent general conditions rather than rare, extreme scenarios.

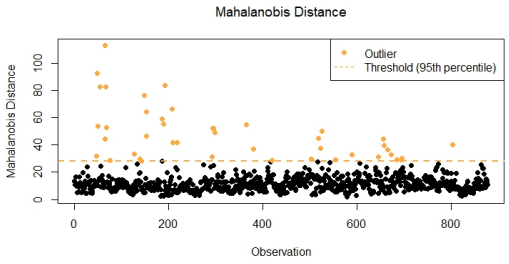

Figure 33: Mahalanobis Distance

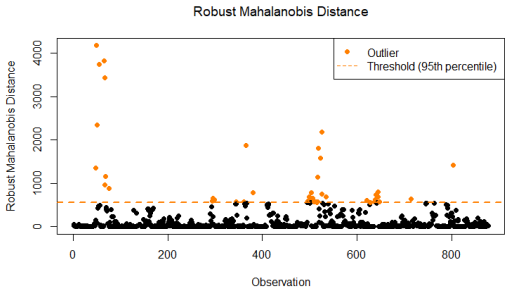

Figure 34: Robust Distance

The two distances can be visualized together in the following Distance-Distance plot (Figure 35).

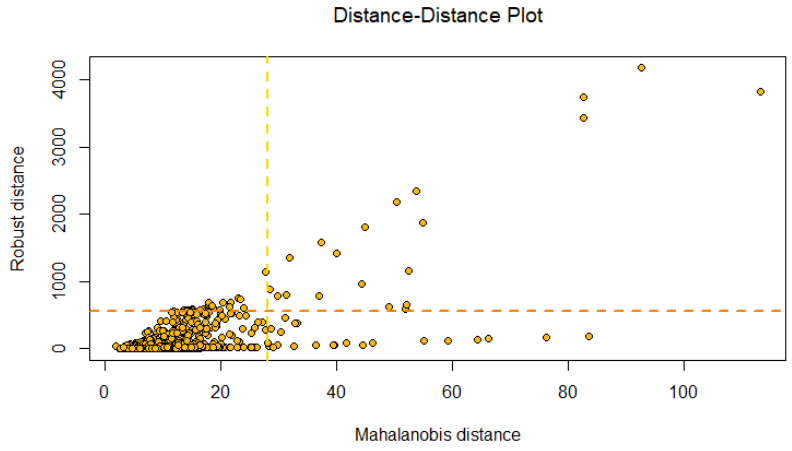

Figure 35: Distance-Distance Plot - After extreme outliers detection

5.3 Best subset selection

After refining the dataset, the next step involves identifying the most meaningful subset of predictors. This ensures that we capture the key drivers of productivity without overcomplicating the model. To this end, a forward selection approach was applied, evaluating variables based on both AIC (Akaike Information Criterion) and Mallow's Cp statistic.

In an initial forward search guided by AIC, variables were added step-by-step, each time selecting the predictor that most reduced the AIC value. The process is illustrated by the results in Table 6:

Step	Predictor Added	AIC Value
1	Days_PM2.5	1786.450
2	nAllNeonic_km2	1775.354

Table 6: AIC Stepwise results

At this stage, the AIC-based approach suggested a parsimonious model that focuses on just these two predictors: Days_PM2.5 and nAllNeonic_km2.

To gain a fuller perspective, Mallow's Cp criterion was also employed. As shown in Table 7 and Figure 36, this method added variables in the following order:

Step	Predictor Added	Cp Value
1	Days_PM2.5	-9.434341
2	nAllNeonic_km2	-28.052785
3	Days_NO2	-34.459930
4	palmer_index	-36.414687
5	Days_CO	-36.822572

Table 7: Cp Stepwise results

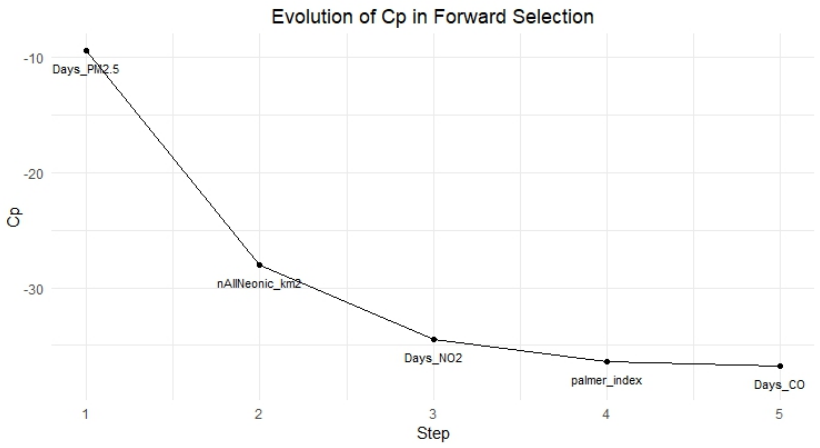

Figure 36: Evolution of CP in forward selection

Each addition slightly improved the Cp value, although the incremental gains became smaller with each new variable. This pattern suggests diminishing returns as the model grows more complex. While AIC favored a simple, two-variable solution, Cp encourages considering a few more predictors to potentially capture more nuanced aspects of pollution and air quality. For this reason we decided to capture a broader picture of environmental stressors and keep the five variables selected by Cp.

Considering this reduced dataset, we now look again at the outliers using the Robust Mahalanobis Distance (Figure 37), which resulted in a more homogeneous dataset, with fewer outliers, allowing for more reliable conclusions.

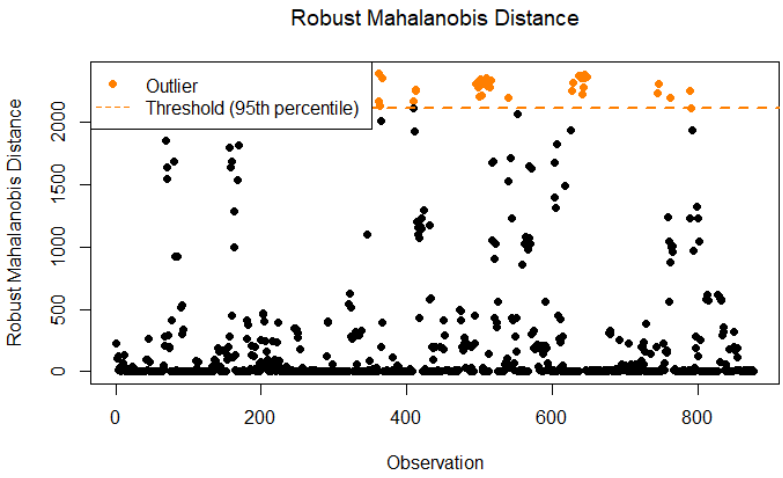

Figure 37: Robust Mahalanobis Distance

6 Mixed Model with Environmental Variables

The goal of this chapter is to explain how external factors influence honeybee productivity, also considering the variability between states due to unidentified reasons, such as environmental policies or differing economic conditions. To achieve this, a mixed-effects model is used that includes the state as a random effect and the variables suggested by the best subset selection process (Table 7) as fixed effects common to all observations.

6.1 Model Evaluation

6.1.1 Q-Q Plot

The Q-Q plot (Figure 38) shows the standardized residuals closely following the theoretical normal line. This suggests that the model's assumption of normally distributed

residuals is reasonable. No substantial deviations from normality are apparent, with just some limited spread on the right side.

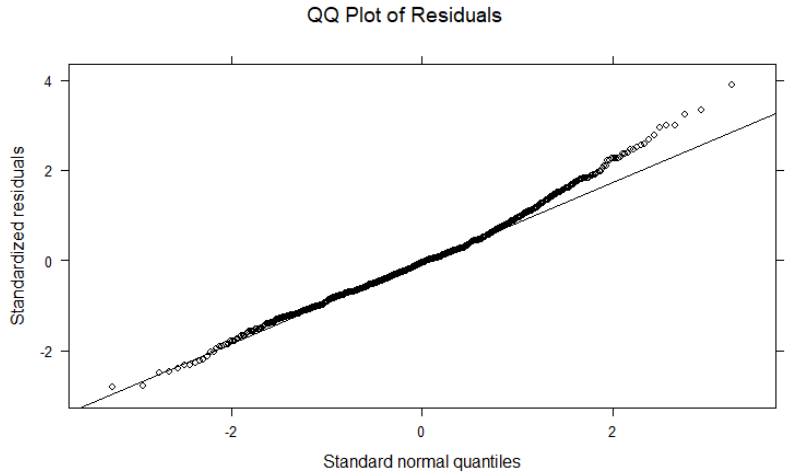

Figure 38: QQ plot

6.1.2 Residual vs. Fitted Values

The residuals (Figure 39) appear randomly scattered around zero with no discernible pattern, indicating that assumptions of homoscedasticity (constant variance) and linearity are largely met.

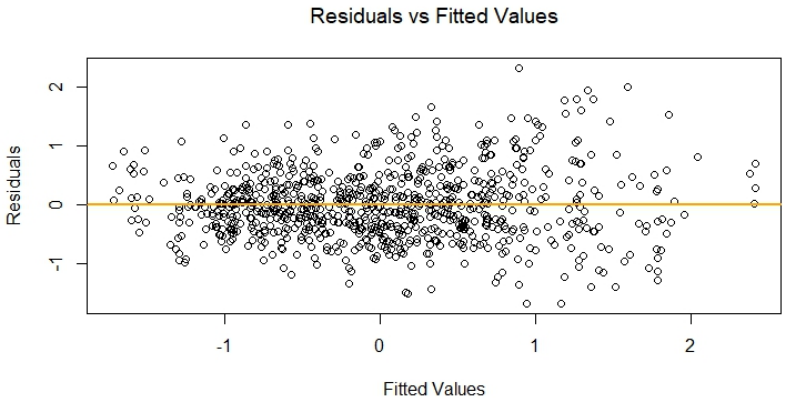

Figure 39: Residual vs Fitted values

6.1.3 Actual vs. Fitted Values

Similarly to the Residual vs. Fitted plot, predicted values align closely with the actual values (Figure 40), clustering around the diagonal line. This alignment suggests the model is reasonably accurate in predicting honey yields.

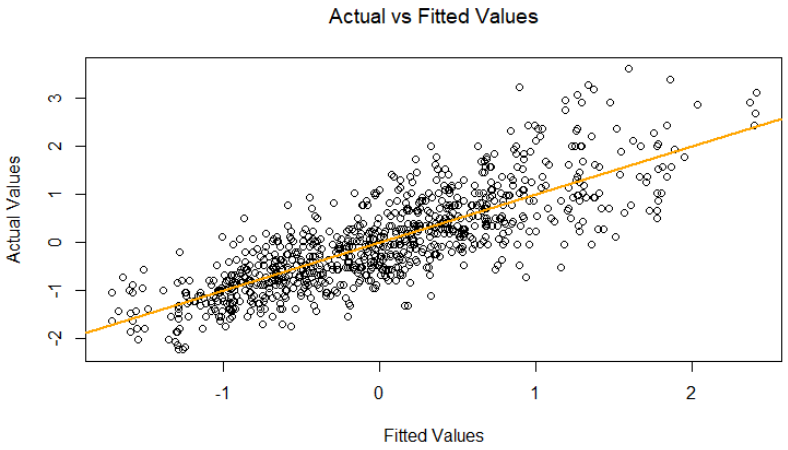

Figure 40: Actual vs Fitted values

6.1.4 Random Effect Dotplot

The dotplot (Figure 41) of random intercepts by state illustrates variability in baseline productivity levels. Some states lie above the global mean, others below, confirming the necessity of a mixed model. This distribution underscores that state-specific conditions, beyond the measured environmental variables, still influence productivity.

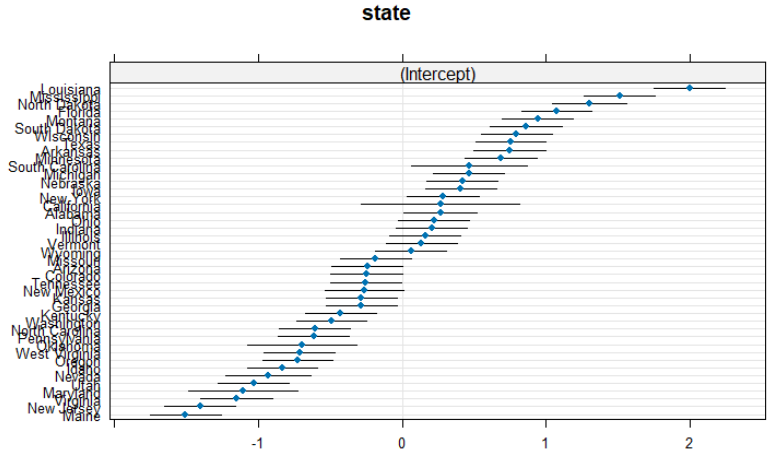

Figure 41: dotplot

6.2 Comparison with Other Models

To validate the mixed model's performance, we compare it with other models: a fixed model and a mixed model with all the variables instead of just the selected ones. This should help understand if the right variables were chosen and if the mixed model with best subset selection is effectively more performing than the others.

6.2.1 R^2 Values

The linear regression without state-level effects has a very low R^2 (0.104), reflecting poor explanatory power. In contrast, the mixed models (with state random effects and environmental covariates) achieve R^2 values around 0.67–0.68, a substantial improvement that confirms the importance of accounting for both state variability and the chosen predictors. The mixed model with just five variables (Mixed Model 2) has a R^2 almost equal to the mixed model that consider all of them (Mixed Model 1), this means that we can safely choose to not include them as they do not improve the model significantly.

Model	R^2 **Value**
Linear Regression	0.104
Mixed Model 1	0.678
Mixed Model 2	0.672

Table 8: R^2 values for Linear Regression and Mixed Models.

6.2.2 RMSE

The Root Mean Square Error (Table 9) for the mixed model (0.583) is markedly lower than that of the simple linear regression (0.946), indicating more accurate predictions.

Model	**RMSE**
Linear Regression (Fixed)	0.946
Mixed Model	0.583

Table 9: RMSE values for Linear Regression and Mixed Model.

6.2.3 AIC Comparison

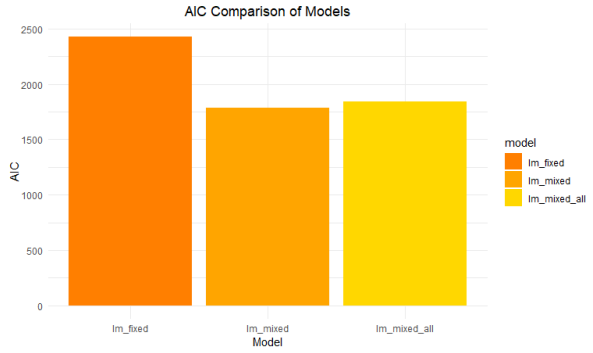

Figure 42: AIC Comparison

The AIC comparison (Figure 42) demonstrates that the mixed-effects model with only five variables provides a better goodness-of-fit compared to the other models. This is because AIC imposes a stronger penalty on overly complex models, leading to a higher AIC value for the model that includes all variables.

6.3 Fixed and Random Effect Summary

Now that the validity of the model has been confirmed, attention shifts to the individual variables. The results of the Linear Mixed Model (Table 10) reveal that Days_PM2.5 and nAllNeonic_km2 have a strong negative impact, indicating that particulate pollution and neonicotinoid pesticides play a detrimental role in honeybee productivity. The palmer_index, which reflects drought severity, suggests that harsher climatic conditions impair productivity, although its effect appears to be relatively modest. Additionally, Days_NO2 and Days_CO exhibit weaker effects; however, their subtle signals should not be overlooked. Previous analyses indicated that their impact might be less pronounced, but they could still carry valuable information when considering the complexities of real-world environmental influences.

Fixed Effects	**Estimate**	**Std. Error**	**t-value**
Intercept	-0.01099	0.12542	-0.088
Days_PM2.5	-0.16634	0.03161	-5.263
nAllNeonic_km2	-0.10050	0.02467	-4.074
Days_NO2	0.07475	0.03562	2.099
palmer_index	-0.04549	0.02246	-2.025
Days_CO	0.04370	0.03166	1.380
Random Effects	**Variance**	**Std. Dev.**	
State (Intercept)	0.6564	0.8102	
Residual	0.3591	0.5992	
Model Summary			
Number of observations		877	
Number of groups (state)		43	
REML criterion		1769	

Table 10: Results of the Linear Mixed Model.

6.4 Bootstrap

It is important to acknowledge the uncertainty inherent in parameter estimates. To gain more robust measures of uncertainty, we employed a bootstrap procedure, which involves repeatedly resampling the dataset with replacement and re-estimating the model parameters. This approach generates distributions for each parameter, allowing us to construct confidence intervals and assess the stability of our findings. Figure 43 shows

the final results. Days_PM2.5 and nAllNeonic_km2 have a robust negative association with honey yield, as expected. Variables like Days_NO2, Days_CO, and Palmer_index may still matter, but their effects are less precisely estimated.

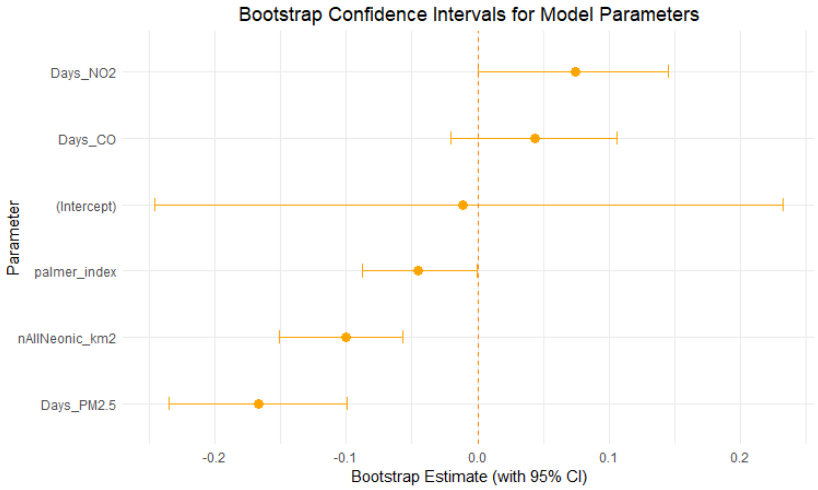

Figure 43: Bootstrap results

7 Conclusion

This study investigates potential connections between honey bee productivity and various environmental and atmospheric factors across the United States. By utilizing a combination of clustering methods, multivariate analyses, and mixed-effects modeling, valuable insights were obtained into the decline of honey production and the regional differences influencing this trend.

Robust clustering methods identified four distinct groups of states in both 1991 and 2021, with these clusters differing in honey price, productivity, and colony density. These findings highlight the diverse conditions and challenges encountered by beekeepers nationwide.

The analysis identified a clear downward trajectory in honey yield per colony over the period examined. Particulate pollution (Days_PM2.5) and neonicotinoid pesticide exposure were consistently associated with reduced honey yields, emerging as significant stressors. Other factors, such as drought severity (Palmer index) and certain pollutants like NO2 and CO, also influenced productivity, though their effects appeared less pronounced. Mixed-effects modeling demonstrated that incorporating state-level variability improved the understanding of these dynamics.

Several limitations were noted during the analysis. Data availability posed challenges, as missing or incomplete records restricted geographic and temporal coverage.

Outliers, such as those originating from California, were addressed in the modeling process but remain partially unexplained.

Additionally, focusing on a subset of environmental variables oversimplified the complexity of factors influencing honey bee productivity. Land use changes, habitat fragmentation, diseases, and beekeeper interventions were not fully represented, limiting the capacity to capture the full spectrum of drivers. While strong correlations between certain stressors and declining yields were identified, causality was not established.

Despite these constraints, the study underscores the considerable influence of pollution and pesticides on honey bee productivity and highlights the importance of addressing localized conditions to better support bee health and sustainability.

Variable name	Description
state	The U.S. state in which the data was recorded.
year	The year of observation.
yealdpercol	Honey production per colony, measured in pounds.
priceperlb	Price of honey per pound.
numcol	Number of bee colonies per state.
area_km2	Land area of the state in square kilometers.
colperkmn2	Density of bee colonies, calculated as the number of colonies divided by the state's area in km^2.
nCLOTHIANIDIN_km2	Clothianidid pesticide use density per km^2.
nIMIDACLOPRID_km2	Imidacloprid pesticide use density per km^2.
nTHIAMETHOXAM_km2	Thiamethoxam pesticide use density per km^2.
nACETAMIPRID_km2	Acetamiprid pesticide use density per km^2.
nTHIACLOPRID_km2	Thiacloprid pesticide use density per km^2.
nAllNeonic_km2	Sum of all the pesticides.
Max_AQI	Max value of Air Quality Index reached.
Days_CO	Number of days with high CO pollution levels.
Days_NO2	Number of days with high NO2 pollution levels.
Days_Ozone	Number of days with high Ozone pollution levels.
Percent_Good_Days	Percentage of days classified as having healthy air quality.
Percent_Unhealthy_Days	Percentage of days classified as having unhealthy air quality.
Days_PM2.5	Number of days with PM2.5 particulate matter exceeding a specific threshold.
Days_PM10	Number of days with PM10 particulate matter exceeding a specific threshold.
Flood	Number of floods.
Fire	Number of fires.
Storm_Group	Number of storms and similar events.
Precipitation	Total precipitation value.
Anomaly_prec	Number of days with anomaly precipitation levels.
palmer_index	A measure of drought severity, where lower values indicate drier conditions.
Max_Temperature	Maximum temperature reached.
Min_Temperature	Minimum temperature reached.

Table 11: Original table