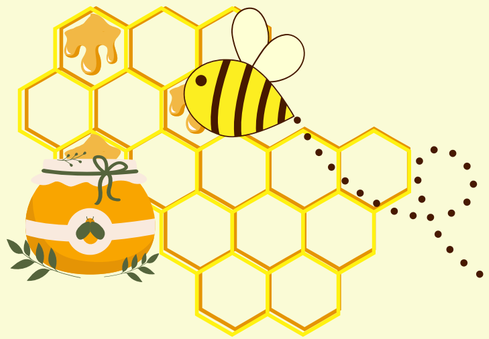
AMS PROJECT

THE IMPACT OF EXTERNAL FACTORS ON HONEY BEES PRODUCTIVITY IN THE USA



Gallo Luigi
Longo Angelica
Rizzi Melissa

WHY HONEY BEES?

Every day, we are reminded of the crucial role that bees play in sustaining life on Earth.
But, bee populations face increasing threats from a variety of environmental and anthropogenic factors, such as climate change, pesticide exposure, habitat loss, and diseases.

Our aim is to explore and quantify the influence of atmospheric and pollution variables on bee productivity.

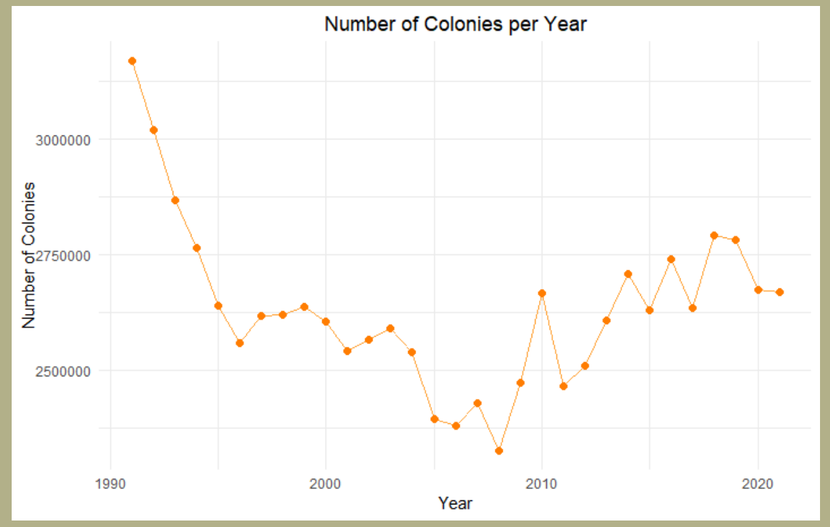
1

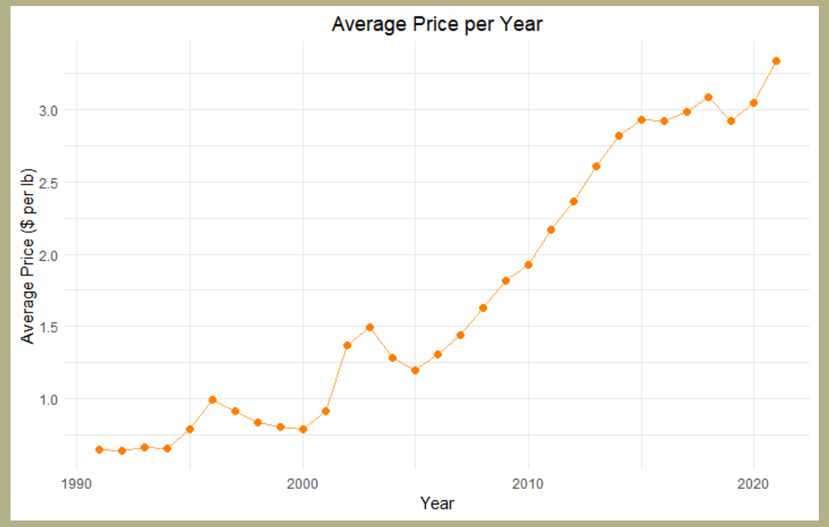
2

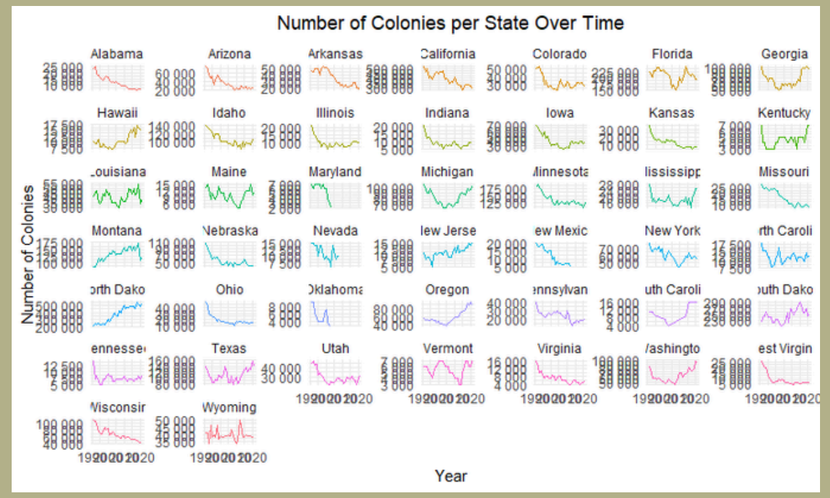
DATA PREPARATION AND DATA VISUALIZATION

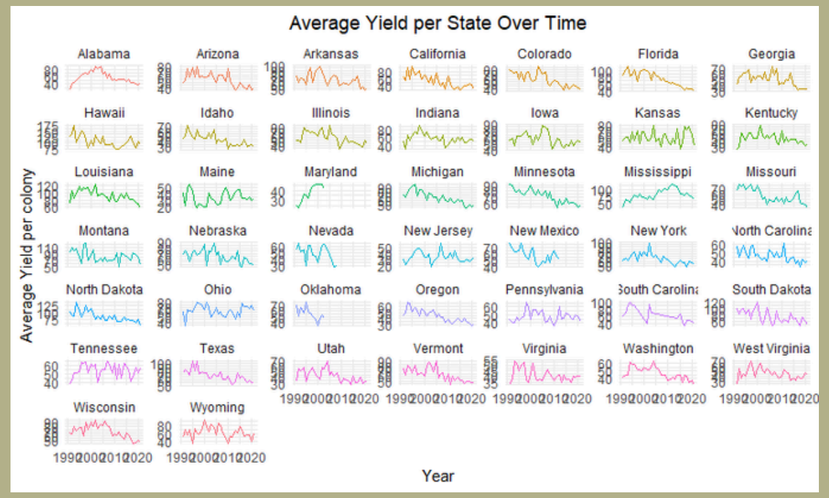
- Ensuring consistent variables names
  e.g.: typos, or inconsistencies
- Conversion of variables recorded in inconsistent units
  e.g.: square miles -> square kilometers, Fahrenheit -> Celsius
- Creation of new variables
  e.g.: colperkm2 = the number of colonies / state area km2,
  nAllNeonic_km2 = sum of all pesticides
- Merging accross different dataset, in order to create a comprehensive dataset grouped by state and year
  variables regarding: bees (colperkm2, yieldpercol, year, state), different type of pesticides used per km2, pollution (Days_NO2, Percent_Good_Days, Days_PM2.5), natural disasters (flood, fire, storm_group), precipitation and temperature (max, min)

EXPLORATORY DATA ANALYSIS









- Considered data: from 1991 to 2021

- **NUMBER OF COLONIES PER YEAR**

A sharp decline can be observed, followed by a more volatile trend in subsequent years

- **NUMBER OF COLONIES PER STATE OVER TIME**
- **AVERAGE YIELD PER STATE OVER TIME**
- **AVERAGE PRICE PER YEAR**

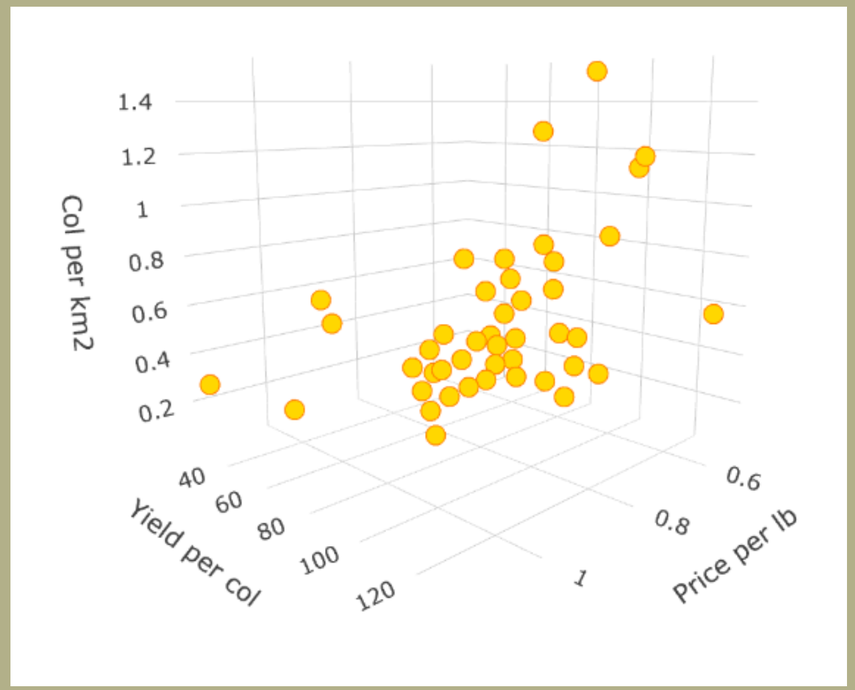
Evidence of inflation and exponential growth over the years

CLUSTERING

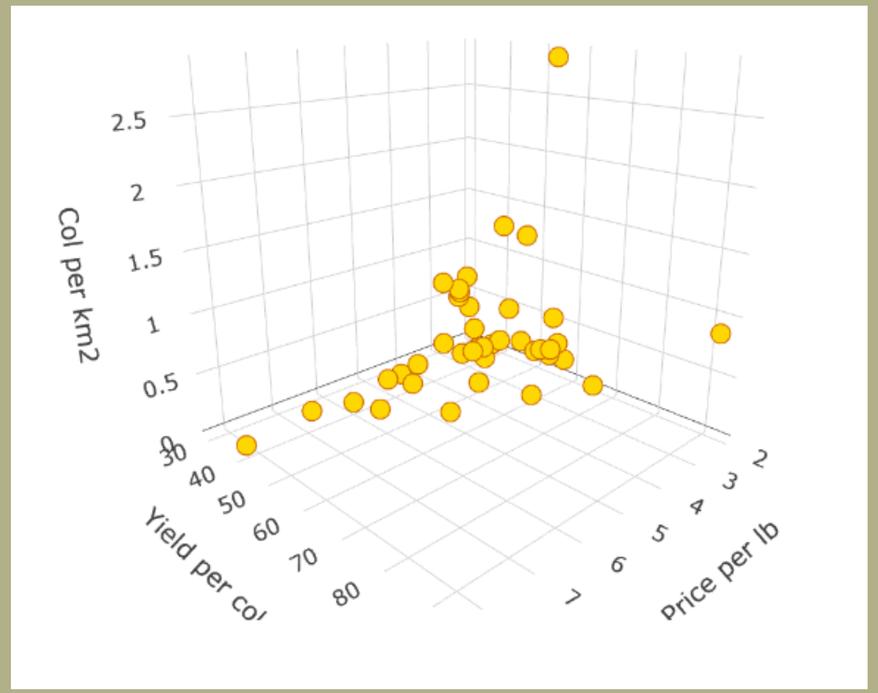
- Aim: identify patterns and similarities between states
- Used method: k-means and robust k-means
- Considered variables: number of colonies per squared kilometer, price per pound of honey, yield per colony
- Comparison of the first (1991) and the last (2021) available year
- Analysis of 40 out of 50 states of USA

CLUSTER - DATA VISUALIZATION

Data visualization - 1991

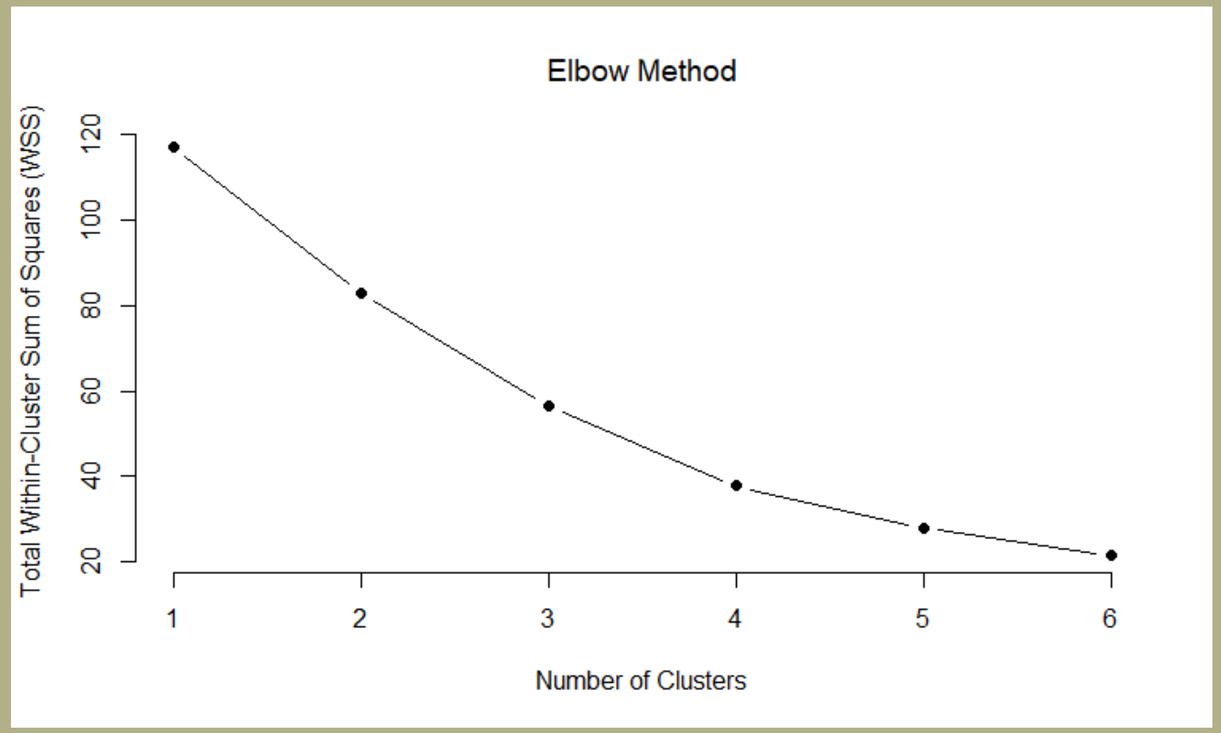


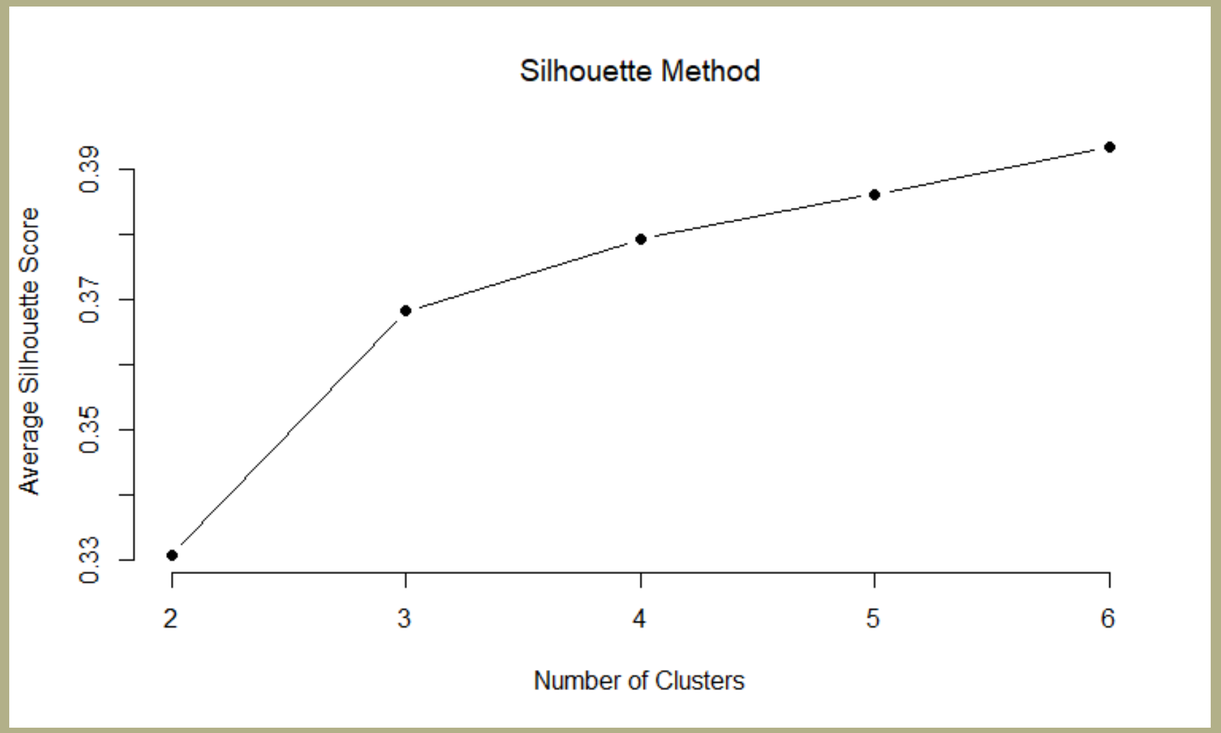
Data visualization - 2021



Presence of outliers can be noticed

K-MEANS - OPTIMAL K (2021)



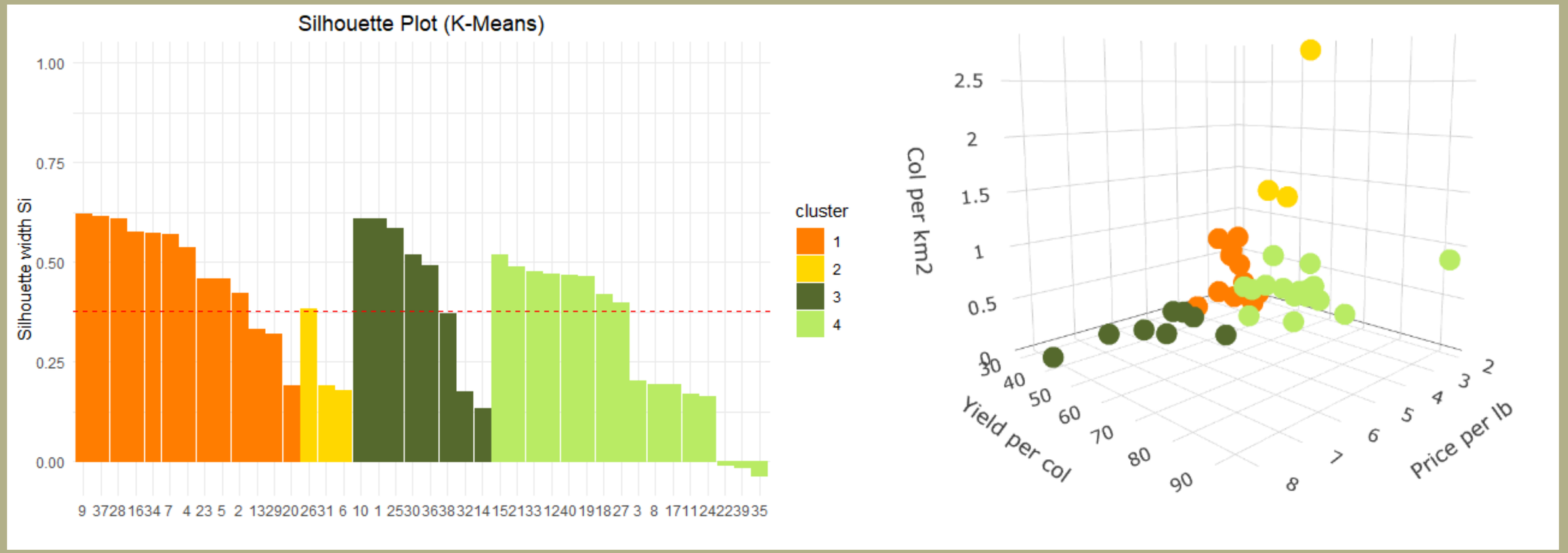


Optimal k: point before the curve flattens

Optimal k: where the average silhouette score is highest

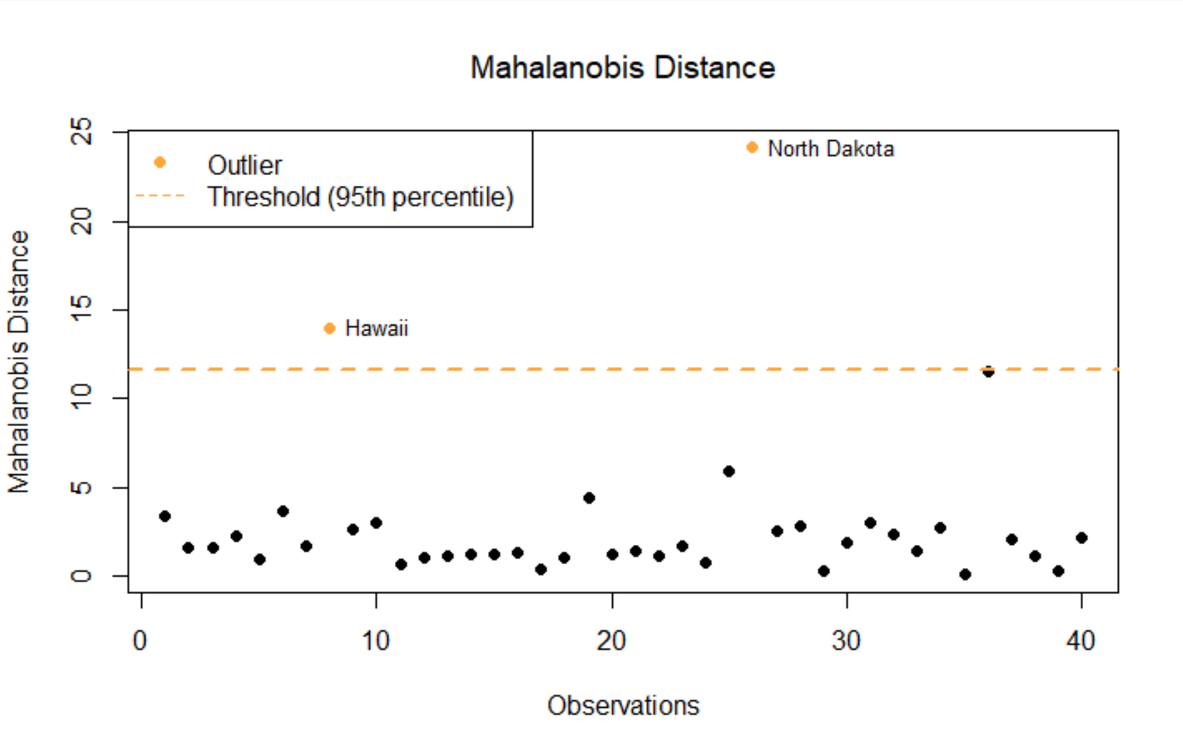
-> Optimal: k = 4

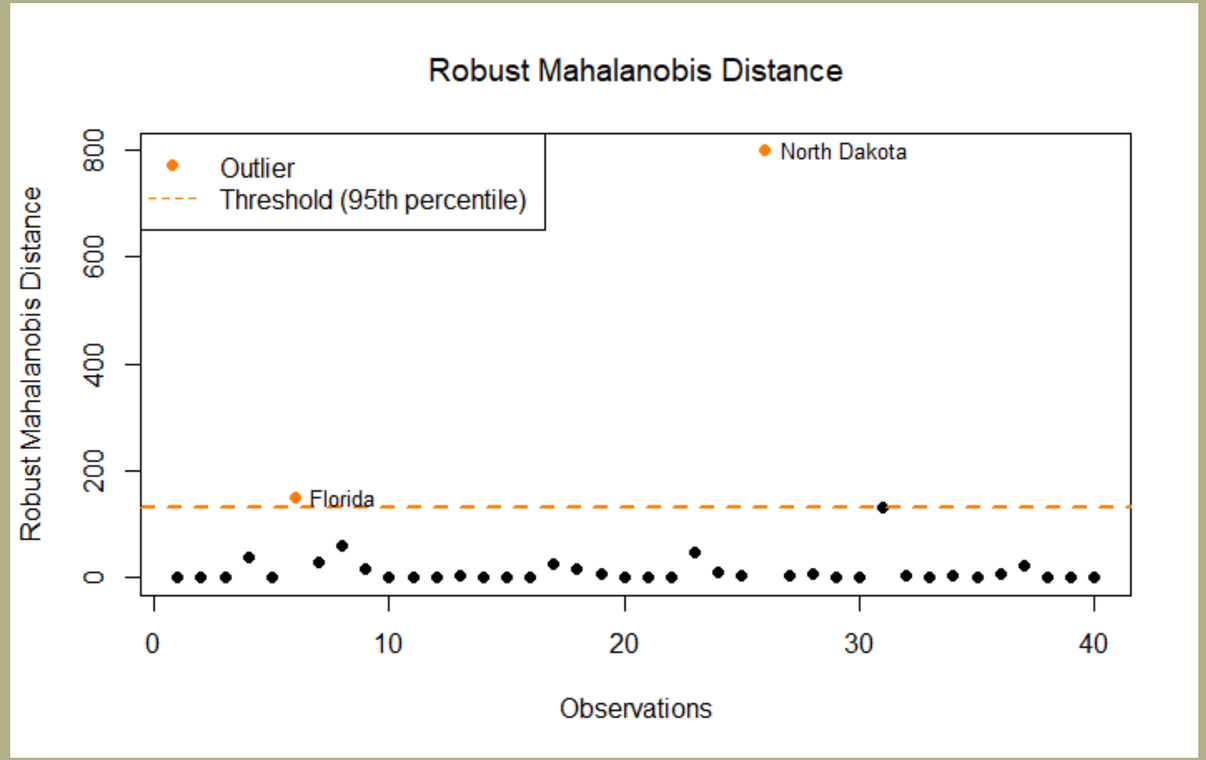
K-MEAN RESULTS (2021)



Evidence of outliers and subsequent misclassified states

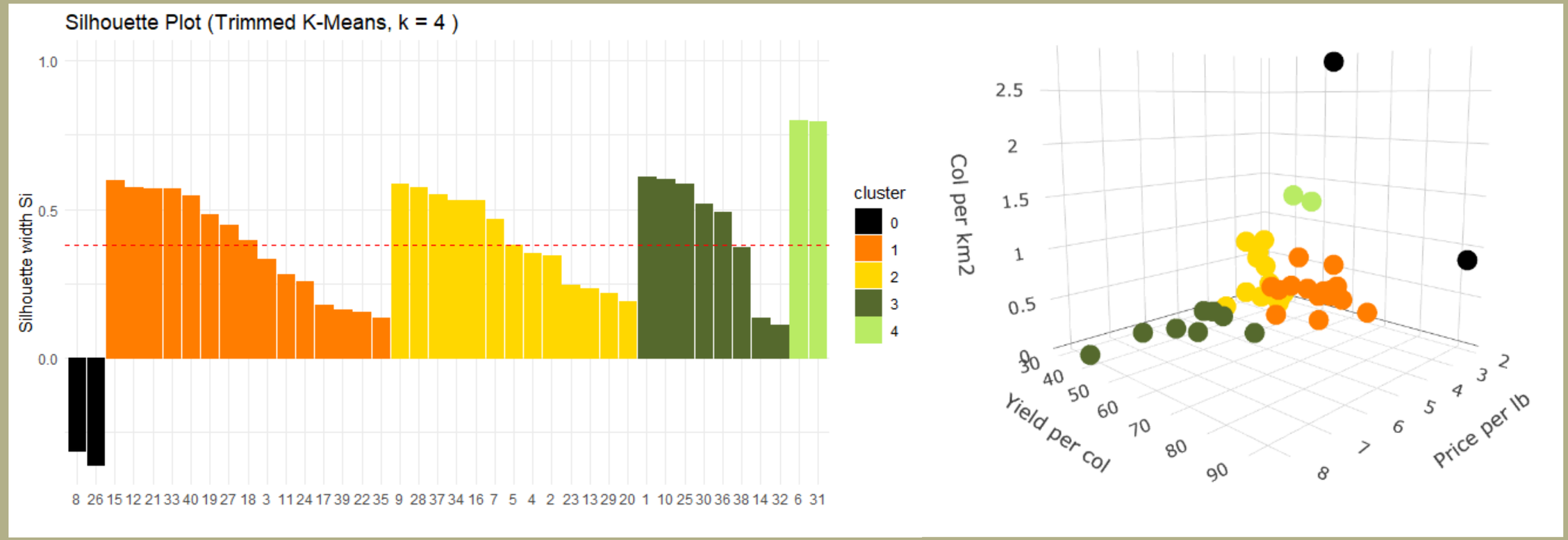
OUTLIERS DETECTION (2021)





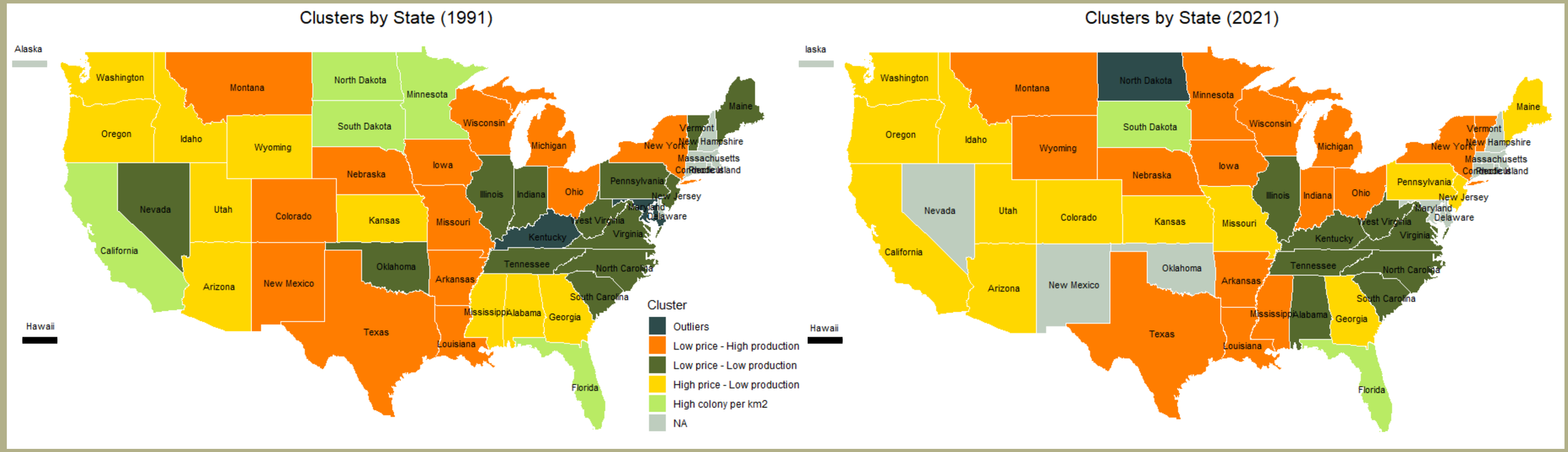
- North Dakota is considered outlier in both cases
- Hawaii is considered outlier in Mahalanobis Distance
- Florida is considered outlier in Robust Mahalanobis Distance

ROBUST CLUSTERING

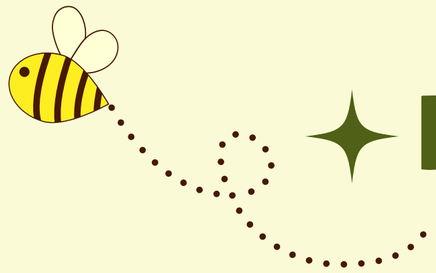


- Dark green: Low productivity - High price
- Orange: High productivity - Low price
- Light green: High colony per km2
- Yellow: Low productivity - Low price

1991 - 2021 CLUSTER COMPARISON



- Dark green: Low productivity - High price
- Orange: High productivity - Low price
- Outliers
- Light green: High colony per km2
- Yellow: Low productivity - Low price
- Non available data



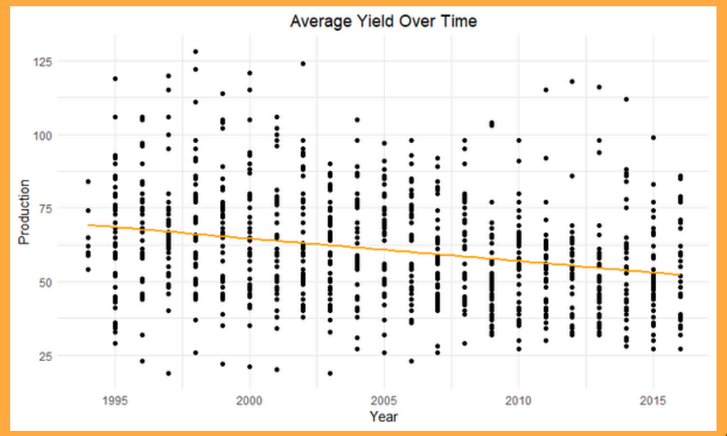
PRODUCTIVITY OVER TIME

- Aim: understand the overall ongoing of productivity over time
- Considered data: from 1994 to 2016

DIFFERENT MODELS

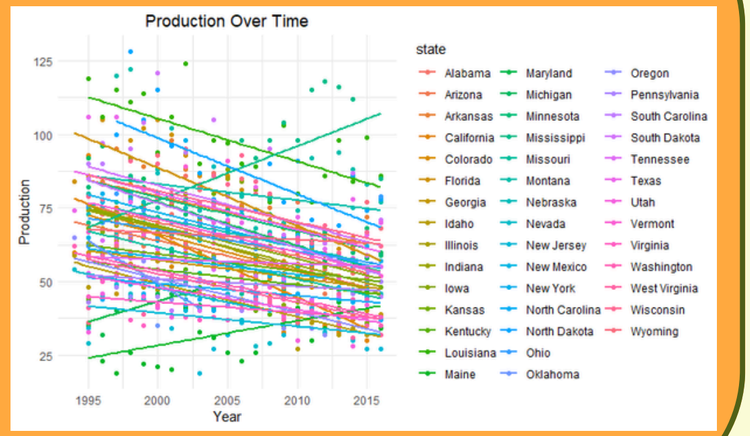
Average productivity over time is analyzed

Full pooling



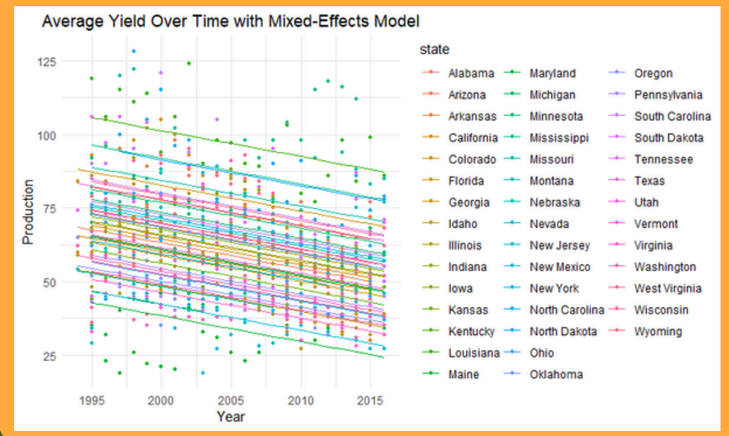
General decreasing trend over time

No pooling



Increasing trend for some states

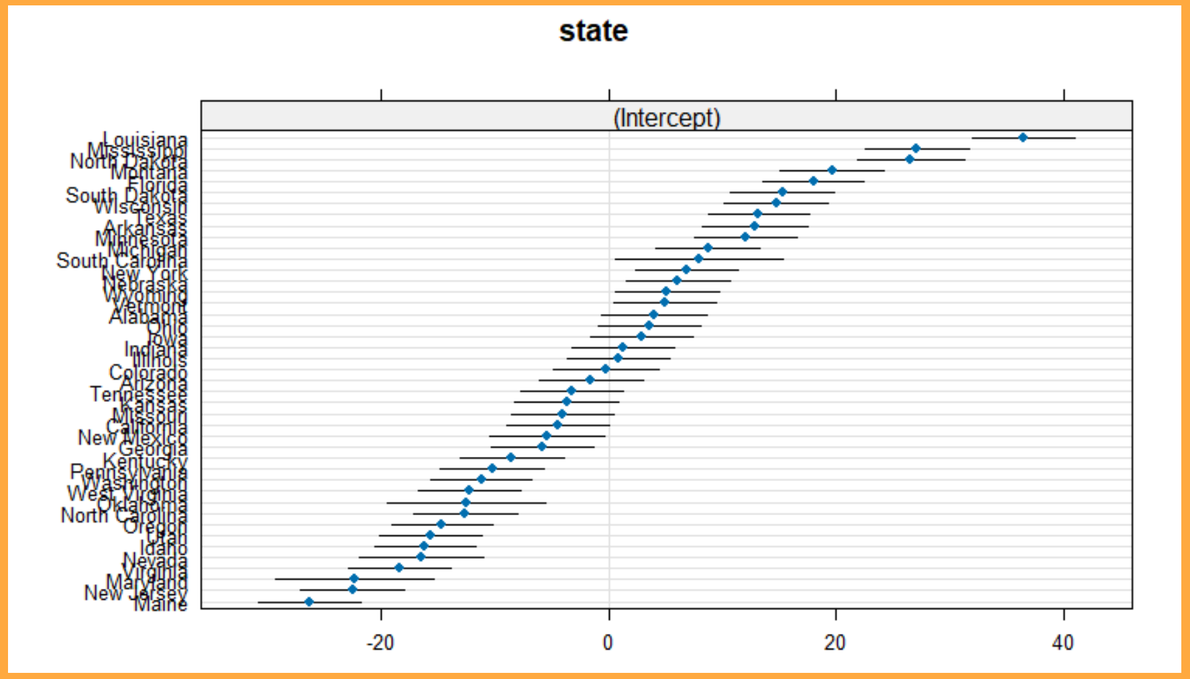
Mixed-effects model



Different intercept for every state

DOES IT MAKE SENSE TO USE A MIXED MODEL?

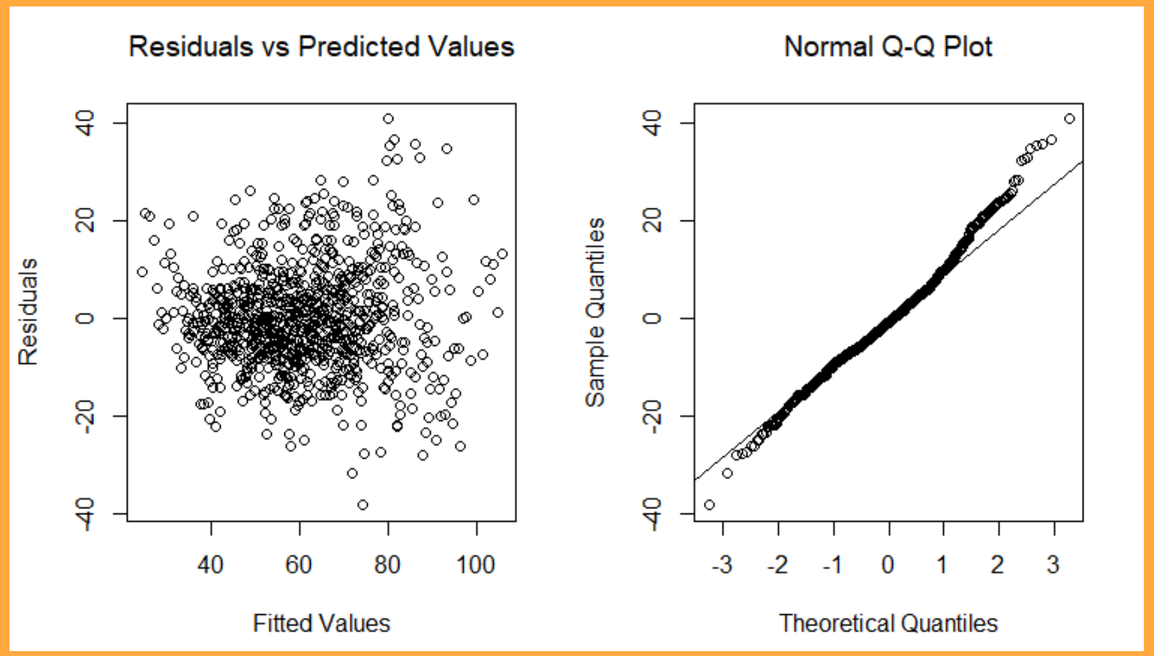
1. Dotplot



The horizontal distribution of points confirms substantial variability in how states deviate from the overall mean

DOES IT MAKE SENSE TO USE A MIXED MODEL?

2. Consistency of residuals with model assumptions



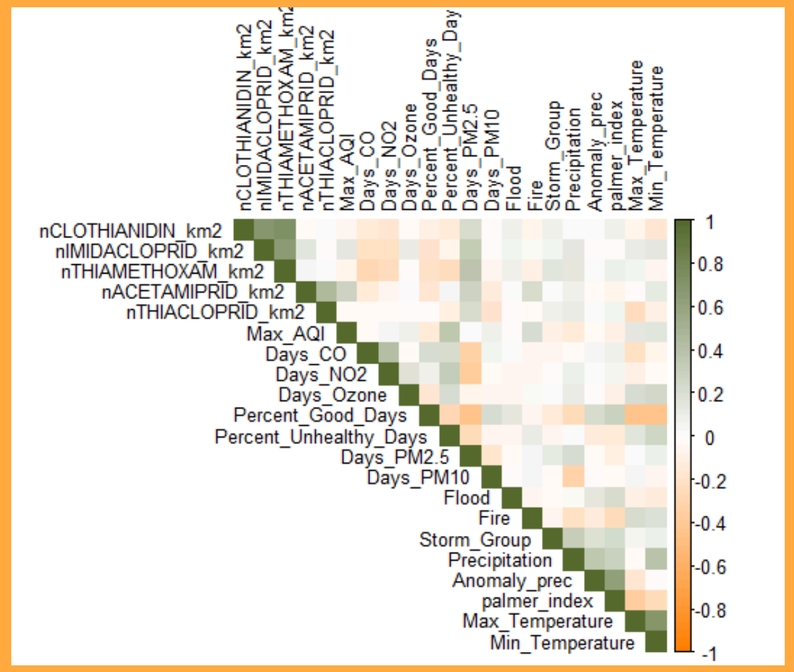
Model assumption: normality of residuals, homoscedasticity of residuals, linearity of fixed effects

VARIABLES ANALYSIS

- Aim: understand which variables should be removed due to correlation, detect outliers and decide how to deal with them
- Used methods: correlation and multiple correlation (VIF) analysis, outliers detection with Mahalanobis Distance and Robust Mahalanobis Distance

CORRELATION AND VIF

Heatmap



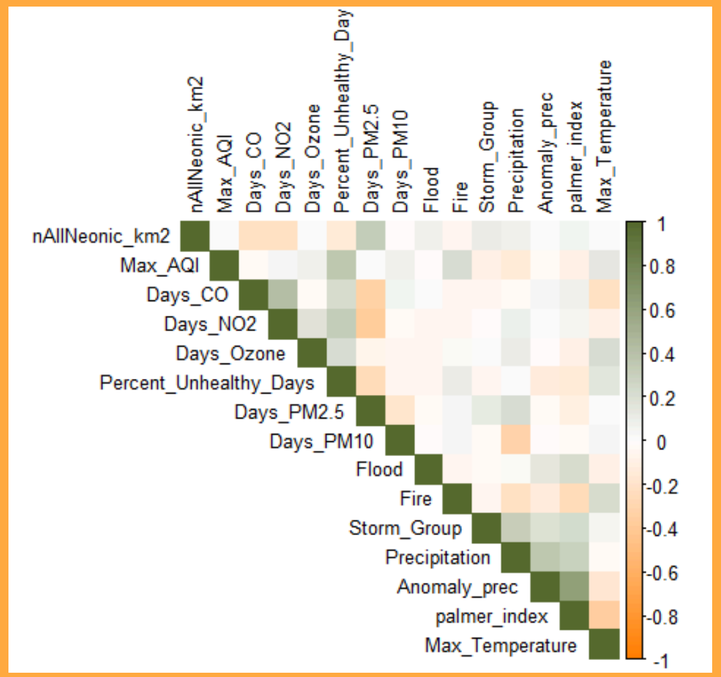
$\sqrt{VIF} > 1,5$

nCLOTHIANIDIN_km2	TRUE
nIMIDACLOPRID_km2	TRUE
nTHIAMETHOXAM_km2	TRUE
nACETAMIPRID_km2	FALSE
nTHIACLOPRID_km2	FALSE
Max_AQI	FALSE
Days_CO	FALSE
Days_NO2	FALSE
Days_Ozone	FALSE
Percent_Good_Days	TRUE
Percent_Unhealthy_Days	FALSE
Days_PM2.5	FALSE
Days_PM10	FALSE
Flood	FALSE
Fire	FALSE
Storm_Group	FALSE
Precipitation	TRUE
Anomaly_prec	FALSE
palmer_index	FALSE
Max_Temperature	TRUE
Min_Temperature	TRUE

Correlated variables: pesticides, max and min temperature,
percent_good_days and unhealthy_days

REMOVAL OF CORRELATED VARIABLES

Heatmap

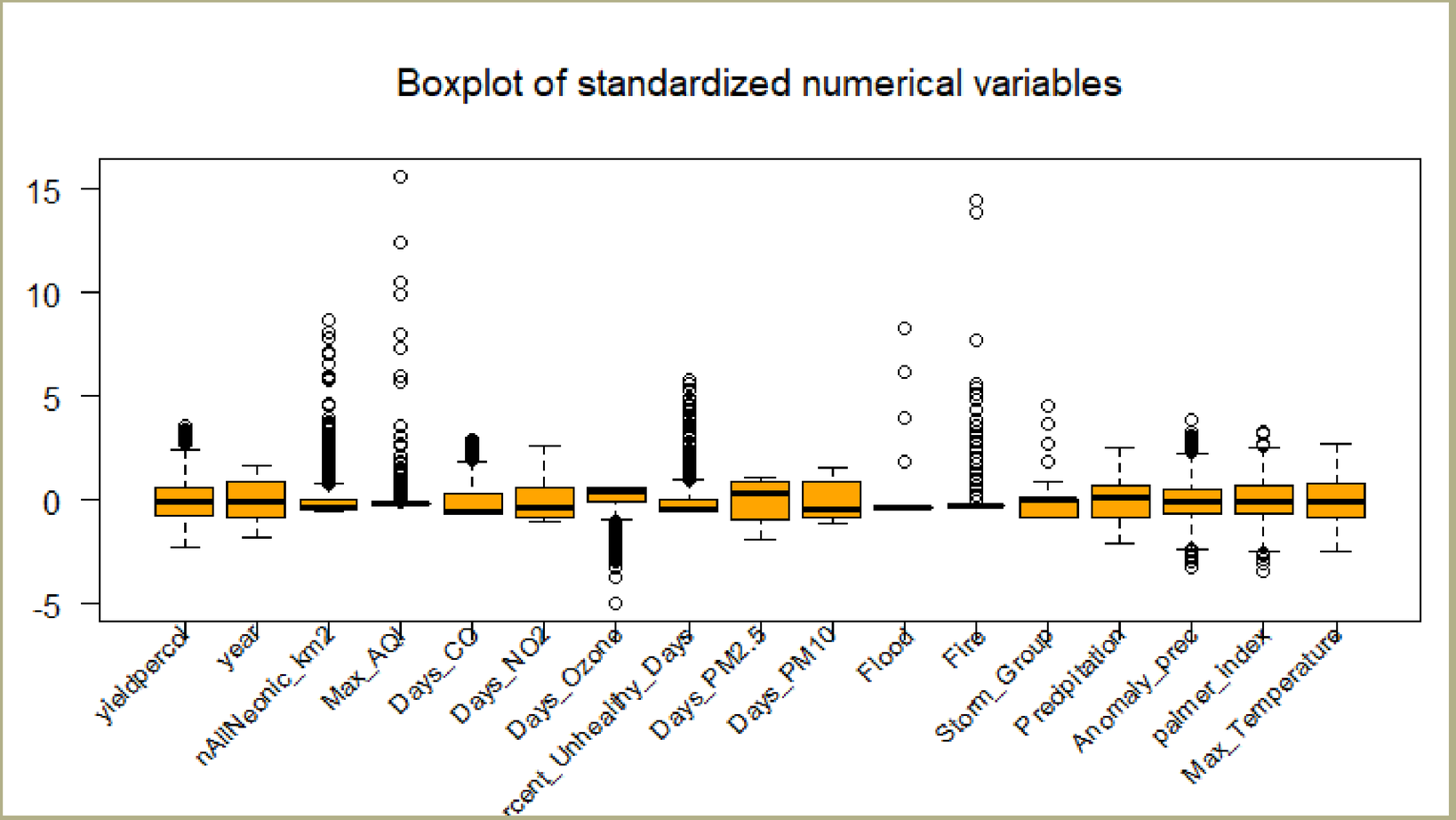


$\sqrt{VIF} > 1,5$

nAllNeonic_km2	FALSE
Max_AQI	FALSE
Days_CO	FALSE
Days_NO2	FALSE
Days_Ozone	FALSE
Percent_Unhealthy_Days	FALSE
Days_PM2.5	FALSE
Days_PM10	FALSE
Flood	FALSE
Fire	FALSE
Storm_Group	FALSE
Precipitation	FALSE
Anomaly_prec	FALSE
palmer_index	FALSE
Max_Temperature	FALSE

Improvement: no multicollinearity

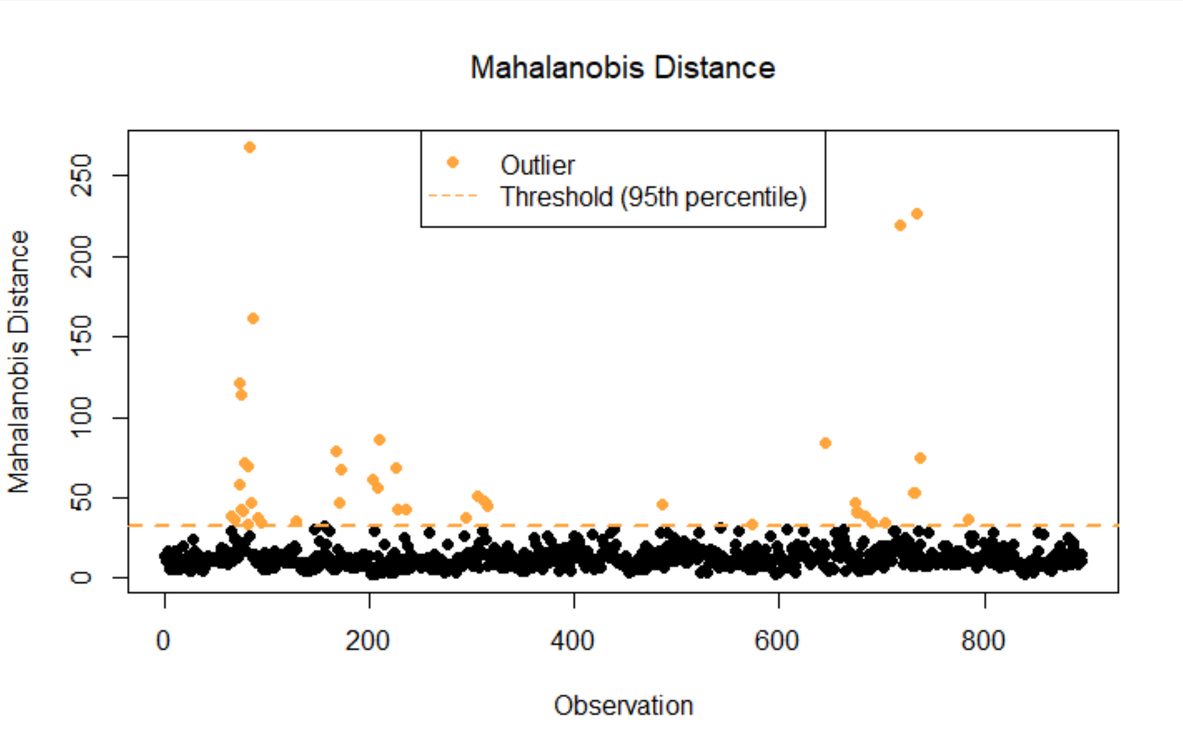
OUTLIER DETECTION

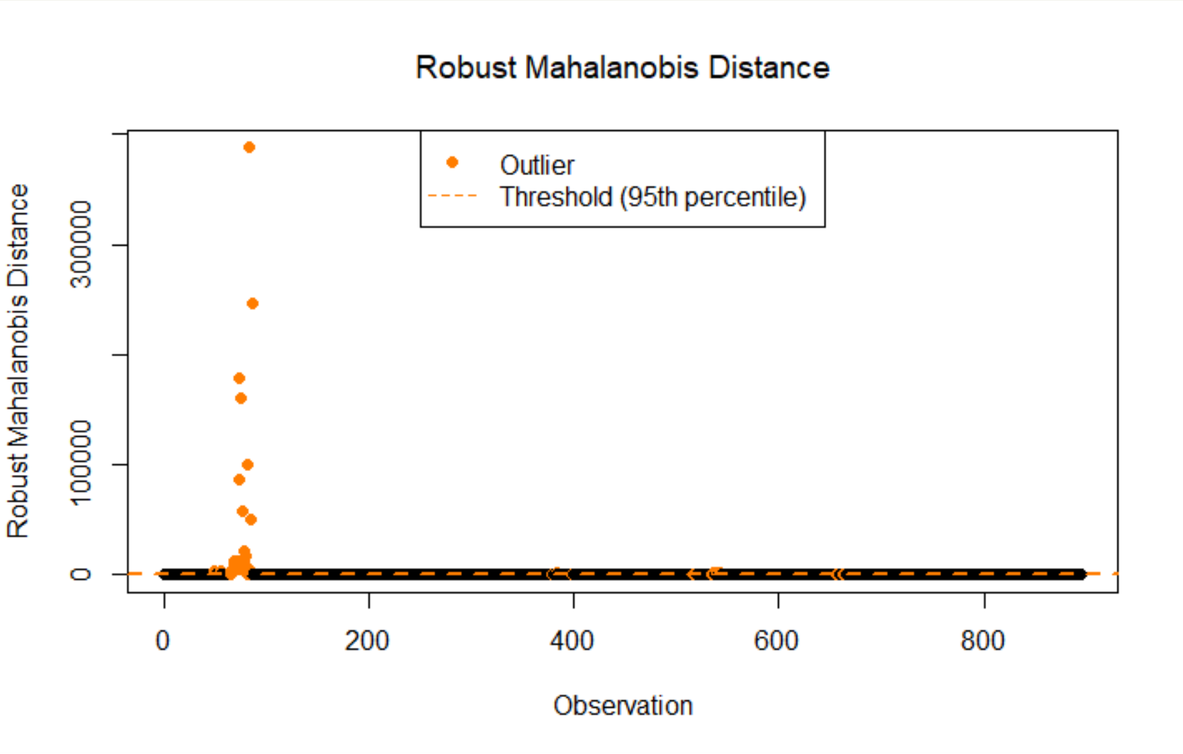


PRESENCE OF OUTLIERS

- Boxplot of standardized variables
- There is evidence of a big amount of extreme values in the single variables
- Are these values considered outliers for the whole dataset?

MAHALANOBIS DISTANCES





- Presence of extreme outliers
- Number of outliers: 45

OUTLIERS REMOVAL

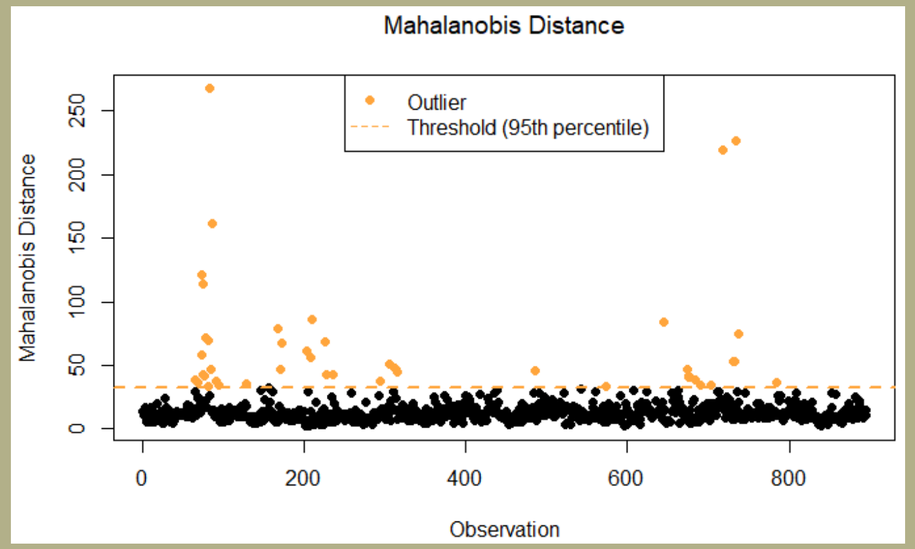
- Extreme outliers (above 98th percentile) removed

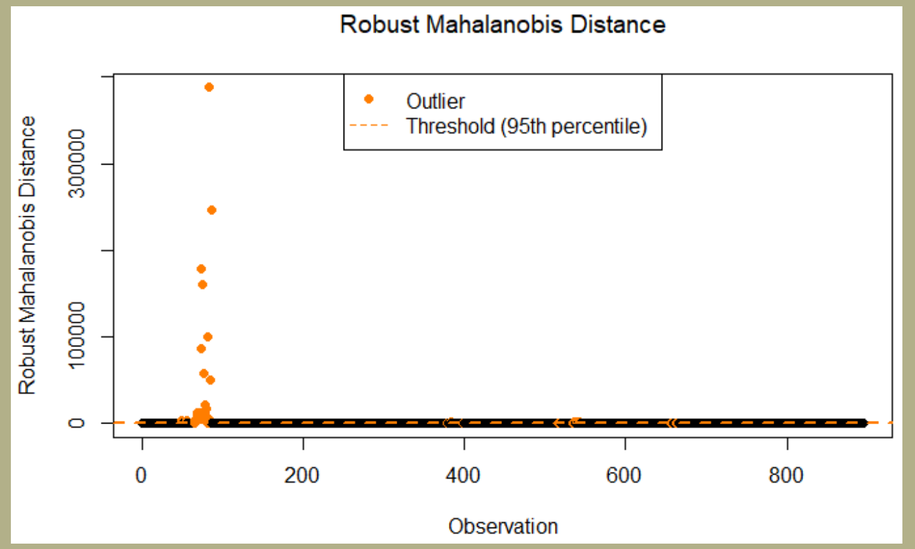
	yieldpercol	year	state	nAllNeonic_km2	Max_AQI	Days_CO	Days_NO2	Fire
7	33	2009	California	0.21580077	15556	7	60	15
8	51	2008	California	0.24898076	2907	12	105	22
9	48	2011	California	0.36508409	14703	8	54	4
10	52	2006	California	0.15745233	9055	6	114	9
11	45	2004	California	0.14483673	5637	13	154	31
12	75	2005	California	0.15143068	4262	14	116	4
13	67	2010	California	0.27185536	4909	5	52	7
14	70	2000	California	0.07519187	11880	77	187	0
15	61	2001	California	0.08589108	22897	82	165	0
16	60	1999	California	0.10200619	3054	27	192	1
17	50	2002	California	0.10909578	8628	29	183	12
18	67	2003	California	0.13531142	18304	16	223	7

- Deleted outliers all about California, why? High values of pollution and fires occurences

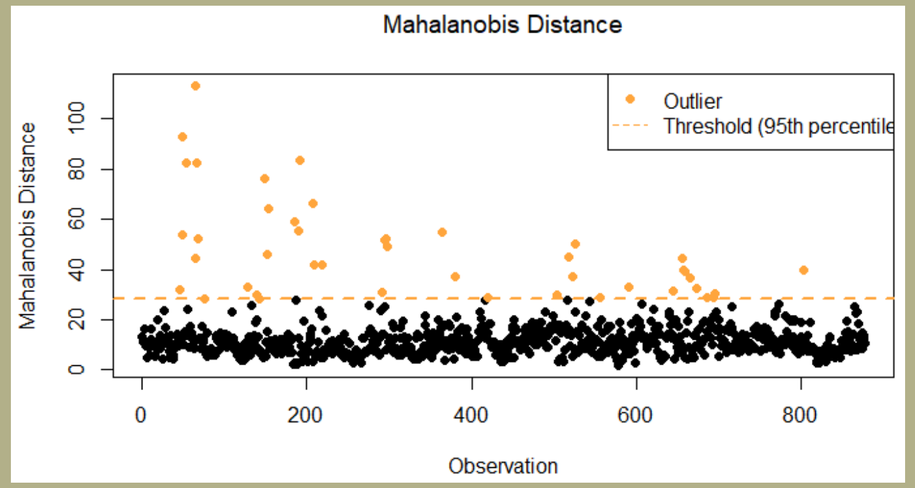
MAHALANOBIS DISTANCES

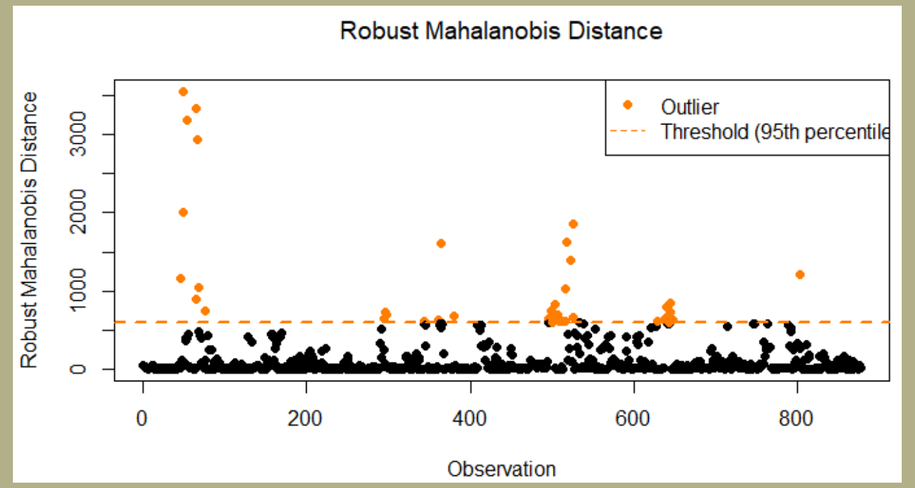
Before outliers removal





After outliers removal





BEST SUBSET SELECTION

- Aim: select variables relevant for the study
- Used method: forward selection with AIC and Mallow CP Index

FORWARD SELECTION

The method

Starts with with an empty model and adds the predictor that most improves the model's performance, until adding more predictors no longer significantly improves the model.

AIC

Focuses on minimizing the AIC, balancing model fit and complexity.

Selected predictors: Days_PM2.5, nAllNeonic_km2

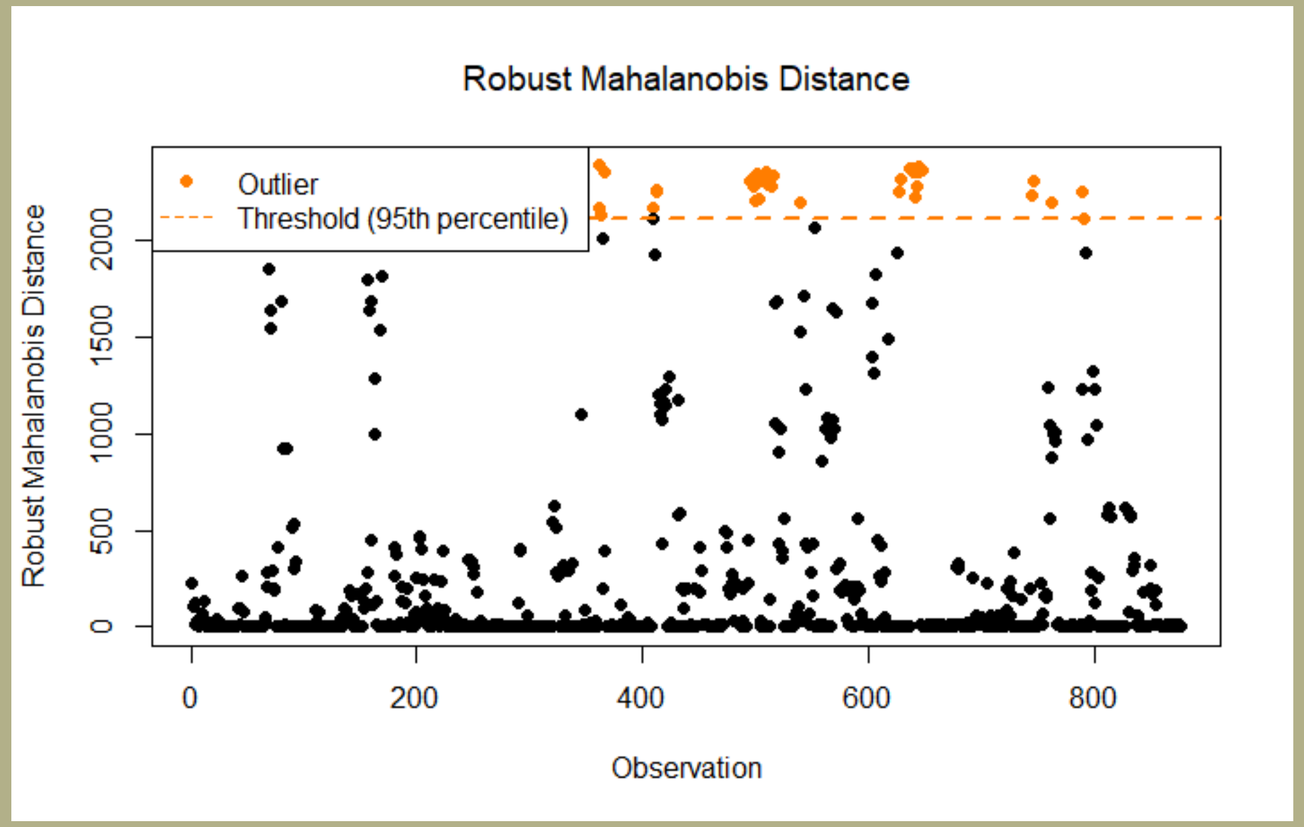
Mallow CP Index

Compares the model's fit to the full model, aiming to minimize prediction error.

Selected predictors: Days_PM2.5, Days_NO2, nAllNeonic_km2, Days_CO, palmer_index

NEW MODEL

$$Y=\beta 0+\beta 1 \cdot Days\_PM2.5+\beta 2 \cdot nAllNeonic\_km2+\beta 3 \cdot Days\_NO2+\beta 4 \cdot palmer\_index+\beta 5 \cdot Days\_CO+\epsilon$$



-> Further improvement

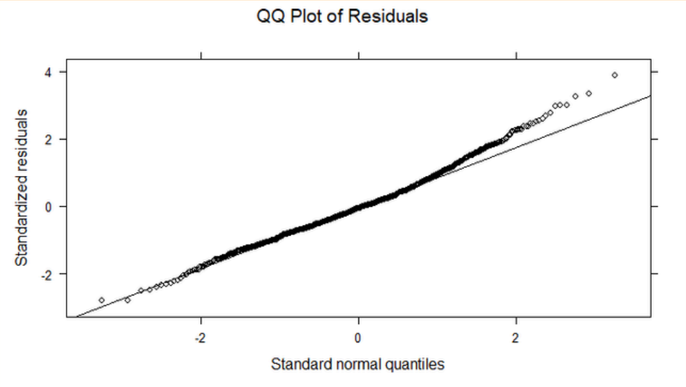
MIXED EFFECT MODELS

Now that we have analyzed the dataset, selected the model, studied the correlation between variables and choosen the relevant ones, we can apply the mixed model and finally understand if there's any relationship between external factors and honey bees productivity.

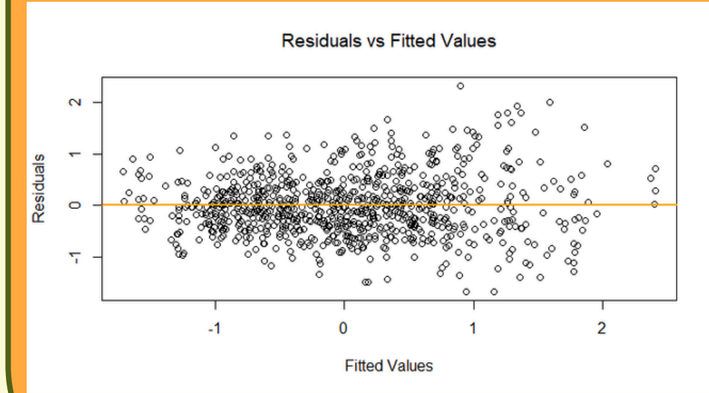
MODEL EVALUATION

Predicted vs Actual

QQ plot of residuals



Residuals vs Fitted



To evaluate the adequacy of the model fit

MODEL EVALUATION

R squared

- Adjust R2 fixed model: 0.089
- R2 mixed effect model with BSS: 0.672
- R2 mixed effect model with all variables: 0.678

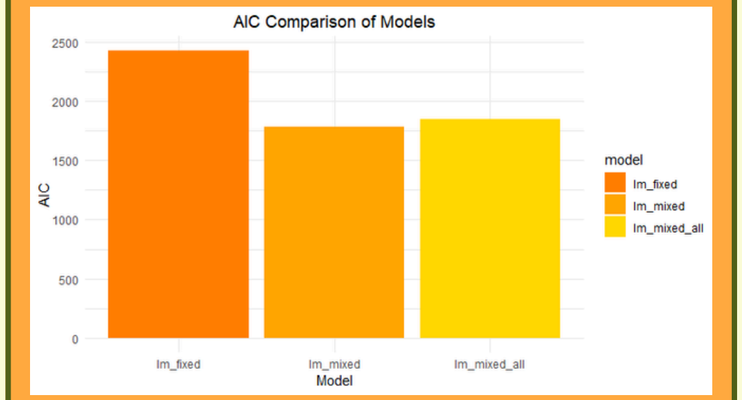
-> Mixed effects model with best subset selection is the best model

RMSE

- RMSE fixed model: 0.946
- RMSE mixed model with BSS: 0.583
- RMSE mixed model with all variables: 0.581

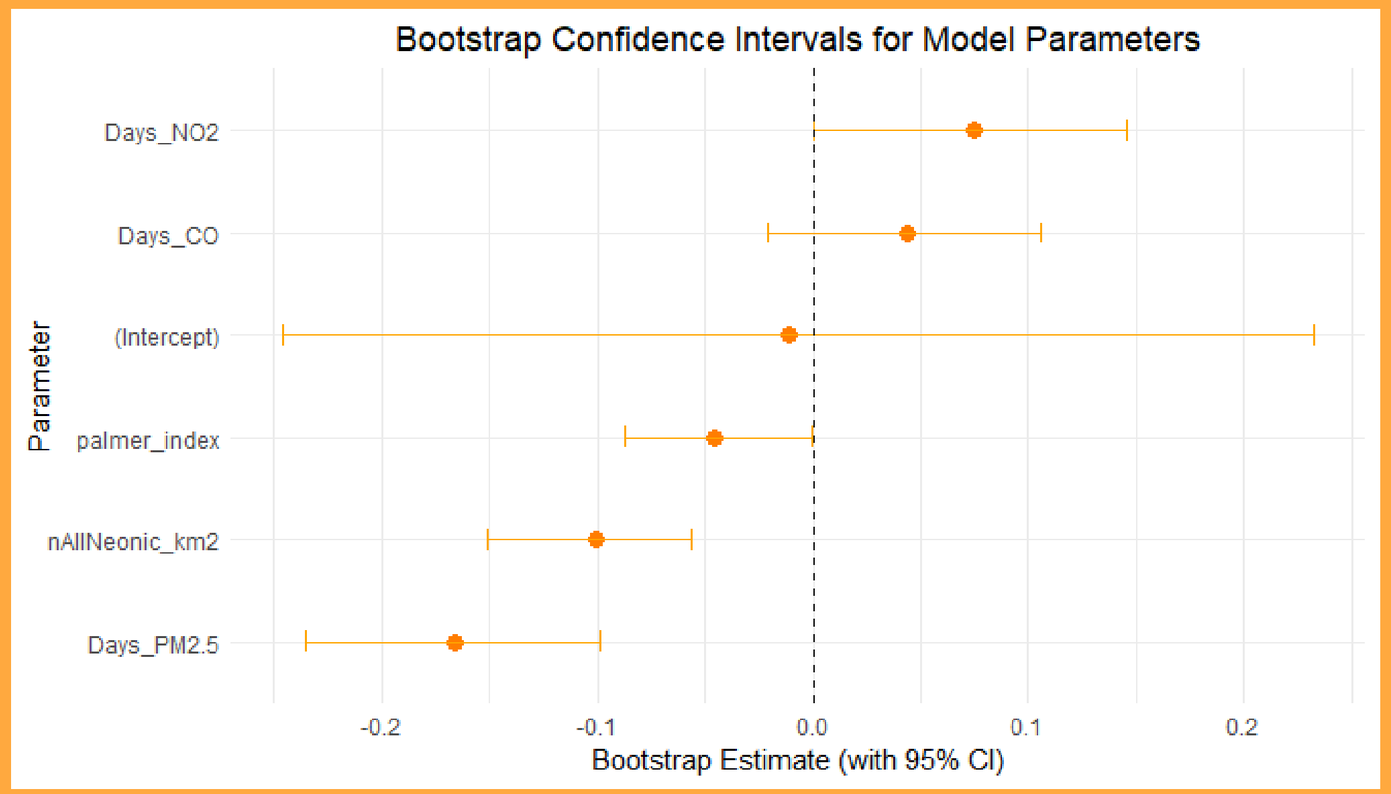
-> Mixed effects model with best subset selection is the best model

AIC comparison



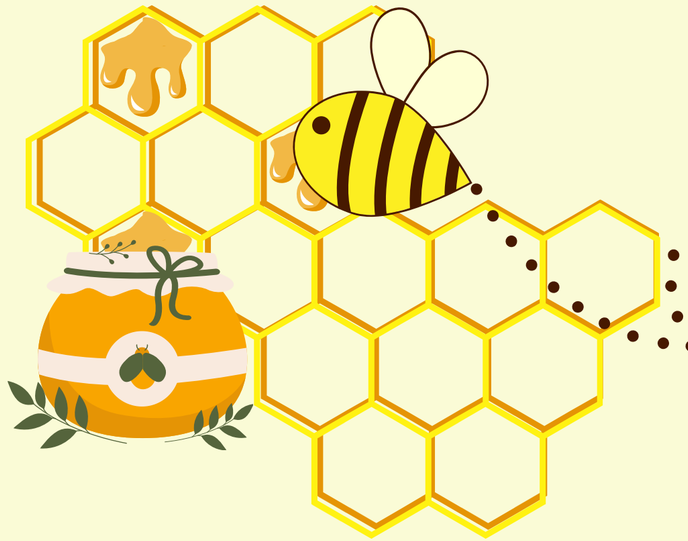
Lower AIC indicates the best model

BOOTSTRAP


Bootstrap Confidence Intervals for Model Parameters
Days_NO2
Days_CO
(Intercept)
palmer_index
nAllNeonic_km2
Days_PM2.5
Parameter
-0.2
-0.1
0.0
0.1
0.2
Bootstrap Estimate (with 95% CI)

PRESENTATION CONCLUSIONS

- Distinct and well-differentiated clusters can be identified, showing a clear geographical division across the states
- A common declining trend in productivity is observed
- This decrease can be largely attributed to two key factors: the usage of pesticides and the number of days with elevated levels of PM2,5



THANK YOU!