Final Report

Forecasting Future Pesticide Usage Trend In US

2242-ECON-5337-001

Business & Economic Forecasting

Spring 2024

Professor: Dr. Christopher Candreva

Shrishankar Shripadarao Desai

Student id-1002173907

1.Introduction

Pesticides, which are either chemical or biological agents, play a critical role in controlling agricultural pests, thus boosting crop yields to support a global population of eight billion. However, their use is not without drawbacks, as they can pose serious health risks and lead to significant diseases. In response, the pesticide industry is diligently working to create more environmentally sustainable solutions. Governmental entities gather essential data from diverse regions, crucial for informing the industry's regulatory and mitigation efforts to minimize both environmental and health impacts. Companies could utilize predictive analytics on this data to project future trends in pesticide usage. This approach would aid in strategic decision-making concerning production scheduling, inventory control, and market strategy.

This project aims to predict the patterns of pesticide use over time for various chemicals in local catchments and watersheds, to assess levels of contamination. It entails forecasting the spatial and temporal distribution of pesticide concentrations, drawing on historical data on usage and environmental conditions.

Assessing the risks posed by pesticides to human health or the environment is a complex and often imprecise process. This complexity stems from various factors, including variations in exposure times and levels, the differing toxicities and persistence of pesticide types, and the unique environmental attributes of the areas where they are applied. Additionally, the number of criteria used, and the methodologies employed to evaluate the negative impacts of pesticides on human health can influence risk assessments. These factors may also impact the evaluation of currently approved pesticides and the approval process for new compounds in the near future.

2.Data Description:

Data link: https://www.kaggle.com/datasets/konradb/pesticide-usage-in-the-united-states

The data on agricultural pesticide use was gathered through the USGS. This data aims to enhance the understanding of pesticide presence in freshwater and its effects on water availability across the United States.

This data consists of six variables namely:

1. **COMPOUND:** This represents the name or type of pesticide compound being used or applied.
2. YEAR: This represents the year in which the pesticide application occurred.
3. **STATE_FIPS_CODE:** This represents the Federal Information Processing Standards (FIPS) code for the state where the pesticide application took place.

4. **COUNTY_FIPS_CODE:** This represents the Federal Information Processing Standards (FIPS) code for the county where the pesticide application took place**.**
5. **EPEST_LOW_KG:** This represents the lower estimate of the amount of pesticide applied, measured in kilograms.
6. **EPEST_HIGH_KG:** This represents the higher estimate of the amount of pesticide applied, also measured in kilograms.

EPEST_HIGH_KG is the primary variable in this study as it represents the higher estimates of agricultural pesticide usage found in water bodies. The dataset comprises 1,048,576 entries, which were cleaned, and missing values addressed using the `na. interp` function. It was then converted into a time series format with annual data points. A logarithmic transformation was applied for normalization, followed by unit root testing to ensure stationarity. The data was subsequently divided into an 80% training set and a 20% testing set. The models employed in the analysis included Simple Exponential Smoothing (SES), Holt's Linear Trend Model, Exponential Smoothing State Space Model (ETS), ARIMA, and Dynamic Regression.

The ARIMA (1,0,1) model was determined to be the most effective based on the RMSE of the back-transformed training data and was used to forecast the next four years.

3.Literature Review:

Several studies have highlighted the health risks associated with pesticide exposure, particularly concerning farmers and other end-users, as well as the general population through residues in food and drinking water. In their article, "Pesticide Exposure, Safety Issues, and Risk Assessment Indicators," Christos A. Damalas and Ilias G. Eleftherohorinos thoroughly discuss the necessity of developing new pesticides with innovative modes of action and enhanced safety profiles. They also explore the potential benefits of alternative cropping systems that rely less on pesticide use, which could reduce exposure levels and mitigate adverse health effects.

Additionally, the article "Are there increasing returns to pollution abatement? Empirical analytics of the Environmental Kuznets Curve in pesticides" by Mikael B. Gustavsson, Andreas Hellohf, and Thomas Backhaus offers an in-depth examination of the risks posed by REACH-registered chemicals to freshwater environments. This study not only compares the environmental hazards of these chemicals with those of five other classes—biocides, personal care products, pesticides, pharmaceuticals, and Water Framework Directive (WFD) priority pollutants—but also evaluates how production volumes and ecotoxicological data impact hazard assessments across various species and taxonomic groups. The findings contribute to a broader understanding of chemical hazard and risk assessment, highlighting the environmental threats posed by industrial chemicals in Europe as reported by ECHA and comparing these risks with those associated with other significant chemical categories. The analysis also explores how production data and ecotoxicological information influence the estimates of Predicted No-Effect

Concentrations (PNEC), enhancing our comprehension of and potential responses to the ecological effects of chemical pollutants.

In the study "Environmental attitudes and drift reduction behavior among commercial pesticide applicators in a U.S. agricultural landscape" by Adam P. Reimer and Linda S. Prokopy, researchers investigated the environmental perspectives, awareness of pesticide drift issues, and the adoption of drift-reduction techniques among commercial pesticide applicators in Indiana. The study categorized the applicators into three groups: those managing industrial weeds (utility right-of-way), those in agriculture, and aerial applicators primarily operating in agricultural contexts.

The findings revealed that while applicators generally held positive environmental attitudes, their concern for pesticide drifting in the areas where they worked was notably low. Despite this, there was a significant uptake in several drift-reduction technologies, with high adoption rates for modifications such as low-drift spray nozzles (88%) and increasing spray droplet size (92%). However, applicators showed less familiarity with specialized equipment like band sprayers, which had only a 13% adoption rate, and methods for identifying sensitive sites like bee colonies and organic farms.

The study also highlighted that among the different types of applicators, those involved in industrial weed management demonstrated the lowest rates of adopting drift-reduction practices. The primary motivations driving the adoption of these practices were the desires to be good neighbors and responsible stewards of the land. The findings suggest that there is considerable potential for implementing more innovative, voluntary measures to enhance awareness of sites that are sensitive to pesticide drift in rural landscapes. This approach could further encourage the adoption of drift-reduction practices among applicators, thereby mitigating the environmental impacts of pesticide use.

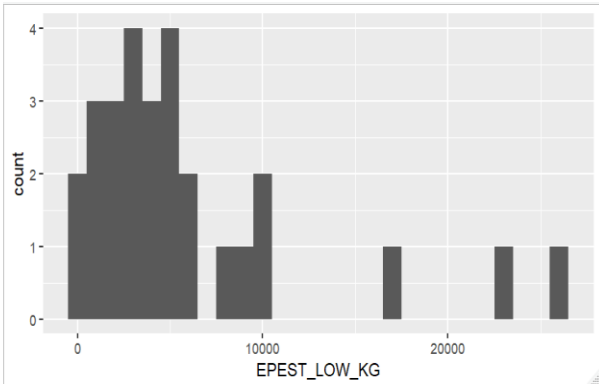
These articles collectively underscore the critical need for regulatory, technological, and methodological advancements in pesticide management and environmental protection strategies. This literature forms the basis of our project's approach to analyzing the implications of pesticide use and exploring viable alternatives and improvements in environmental risk management.

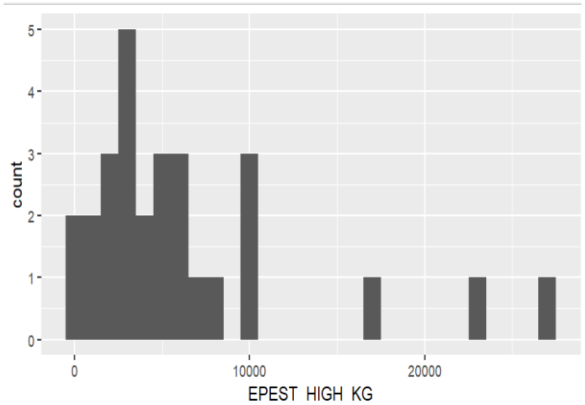
4.Methodology

The initial phase of this project involved data cleaning, where the total count of missing values was determined for each column. Specifically, the column EPEST_LOW_KG had 3,441,641 missing entries, and EPEST_HIGH_KG had 1,276 missing entries. These missing values were addressed using the `na. interp` function to replace the missing values with estimates.

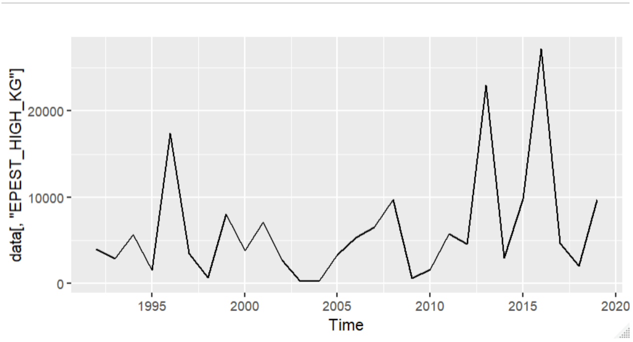
Then used different visual techniques to know the data better. Visual techniques used are histogram, line chart, bar graph, scatter plot.

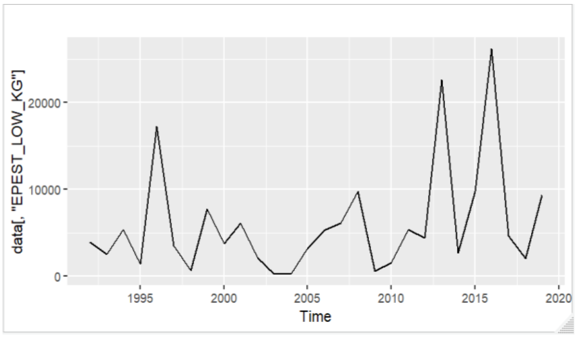
The histogram showcases the distribution of EPEST_HIGH_KG and EPEST_LOW_KG, which represents higher estimates of pesticide usage in kilograms. On the x-axis, these values are segmented into various bins corresponding to different levels of pesticide usage, while the y-axis indicates the frequency of observations within each bin. The data predominantly cluster near zero, indicating that high estimates and low estimates of pesticide usage are generally low. The distribution features multiple peaks, particularly at lower values, which taper off as the values increase, indicating a right-skewed distribution. This skewness suggests that while most instances of pesticide use are minimal, there are occasional higher values.



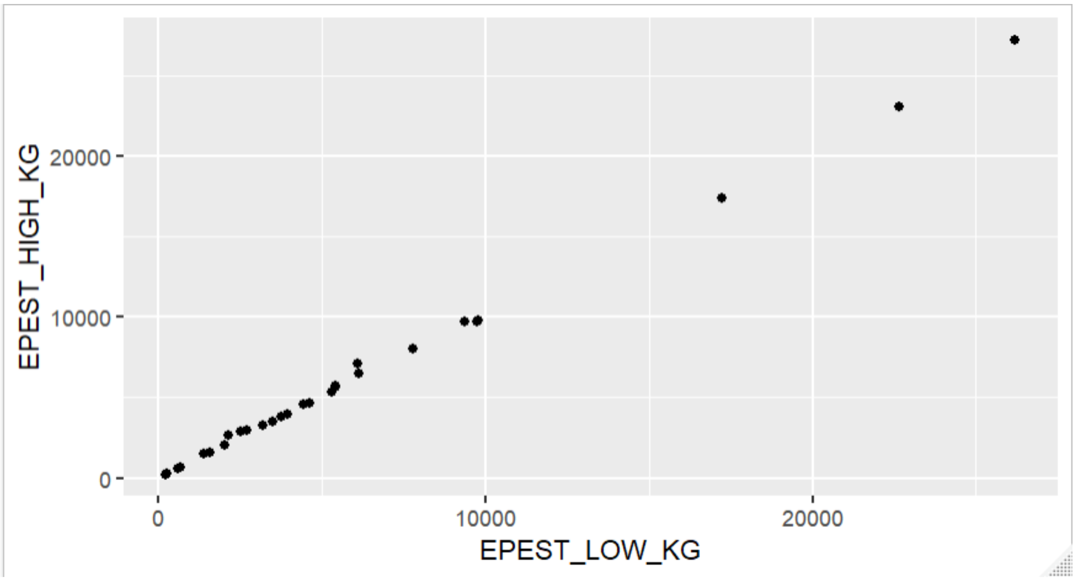


The line graph illustrates the time series of EPEST_HIGH_KG and EPEST_LOW_KG, which quantifies the higher estimates of pesticide usage in kilograms, spanning from the early 1990s to 2020. Throughout this period, the data shows significant variability, with marked peaks approximately every five to ten years. Notably, in the years 2000, 2010, and 2015, pesticide usage surged to levels exceeding 20,000 kg before subsequently declining. This pattern suggests a cyclical nature of pesticide usage, likely influenced by factors such as agricultural cycles, changes in policy, or other external factors that affect pesticide application rates. Despite these fluctuations, there is no discernible long-term upward or downward trend; rather, the data maintains a stable pattern of cyclical peaks and troughs over time.

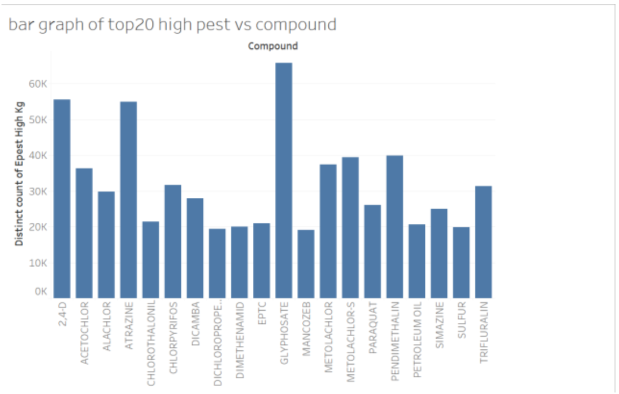




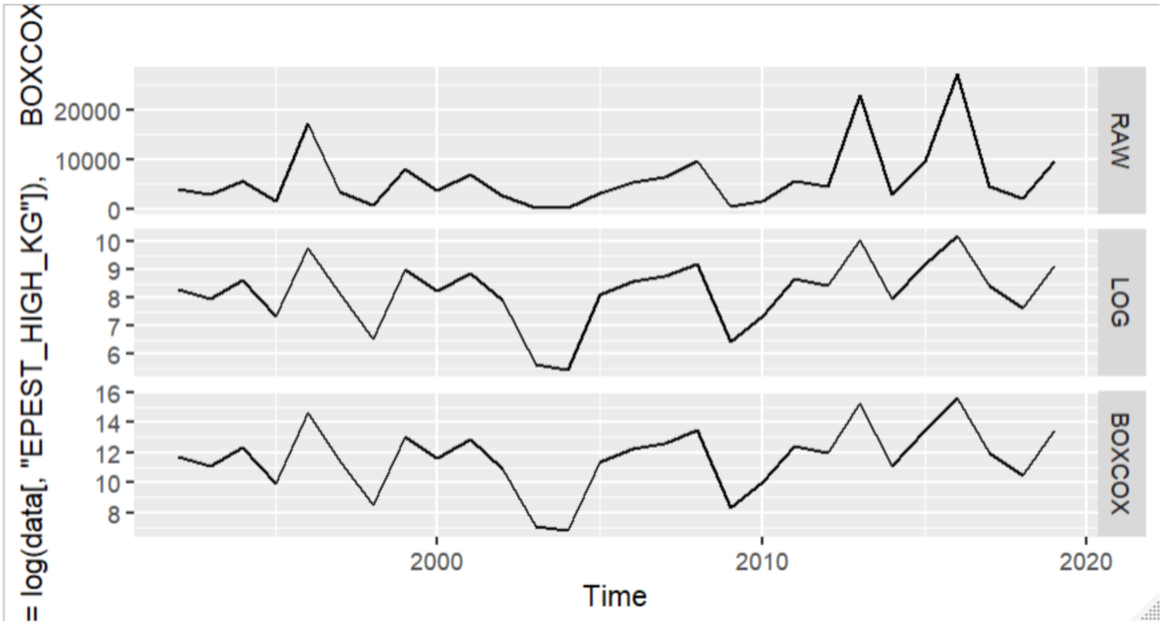
The scatter plot shown here demonstrates the relationship between EPEST_LOW_KG and EPEST_HIGH_KG, which measure the lower and higher estimates of pesticide usage in kilograms, respectively. On the x-axis, EPEST_LOW_KG values span from zero to over 20,000 kg, while the y-axis shows EPEST_HIGH_KG values within the same range. The plot indicates a positive correlation between these two metrics, as higher values of EPEST_LOW_KG tend to be associated with higher values of EPEST_HIGH_KG. Most data points cluster at the lower spectrum of both axes, suggesting that lower pesticide usage estimates prevail more frequently. Nevertheless, several outliers reflect significantly higher estimates, pointing to occasional high levels of pesticide usage.



The bar graph presented visualizes the distribution of the top 20 pesticides by their high estimate usage in kilograms, focusing on distinct counts across various compounds. The graph features a range of pesticides, each represented by a bar indicating the magnitude of its high estimate usage. Notably, certain pesticides show significantly higher usage compared to others, reflecting peaks in the distribution. For example, some compounds reach as high as approximately 60,000 kg, whereas others are noted at lower levels around 10,000 kg to 30,000 kg. This visualization effectively highlights the disparities in pesticide usage, showcasing which chemicals are most prevalently used in higher amounts and providing insights into patterns of pesticide application. Graph obtained in tablue.



A log transformation was utilized to smooth the data, initially characterized by pronounced spikes in pesticide usage. This mathematical adjustment helps stabilize variance and normalize the data distribution, rendering the data patterns more comprehensible and straightforward to analyze. The use of a logarithmic scale moderates the impact of extreme values, thereby improving the graphical representation by minimizing the influence of outliers. This technique not only makes the underlying trends clearer but also enables more precise comparisons across different pesticide compounds, enhancing the understanding of their usage levels. Additionally, the log transformation facilitates compliance with the normality assumptions required by many statistical models, thereby enhancing the reliability of further analytical procedures.



Augmented Dickey-Fuller (ADF) test, applied to the logarithmically transformed EPEST_HIGH_KG data, featuring a drift component. This test is crucial for determining whether the series is stationary or exhibits a unit root, indicating non-stationarity. The key findings include significant residuals that demonstrate the model's fit at each lag, with particularly crucial coefficients for the intercept and the first lag of the series. The intercept is significant, underscoring the importance of the drift component in the model, while the coefficient of the first lag is notably negative and significant, suggesting strong evidence against the unit root hypothesis.

The test statistic value of -14.411, when compared against the critical values for the 1%, 5%, and 10% levels, strongly indicates rejection of the unit root, implying stationarity in the time series data. This is supported by the very high values of Multiple R-squared and Adjusted R-squared, although these should be approached with caution in time series analysis. The F-statistic and its corresponding p-value further affirm the significance of the regression model's terms collectively.

Augmented Dickey-Fuller (ADF) test conducted on the logarithmically transformed data of EPEST_LOW_KG, utilizing a drift component in the model. This test is essential for determining if the time series data exhibits stationarity or contains a unit root, indicating non-stationarity. In the results, the intercept, and the coefficient of the first lag of the series are significantly negative, which is pivotal for the ADF test. A significant negative value of lag1 at a p-value of 0.0316 suggests that the null hypothesis of a unit root can be robustly rejected, indicating the series is stationary.

The test statistic for lag1 is -20.1115, far more negative than the critical values for the 1%, 5%, and 10% significance levels, providing strong evidence against the presence of a unit root. This is supported by the F-statistic and its associated p-value, which further suggest that the regression model's terms are collectively significant.

Models:

Simple Exponential Smoothing (SES)

Three simple exponential smoothing (SES) models were applied to the logarithmically transformed data of EPEST_HIGH_KG. Exponential smoothing models are particularly useful for forecasting time series data by applying smoothing constants to weight observations differently over time. The first model did not specify a smoothing parameter (alpha), and it automatically calculated the best fit, resulting in an Akaike Information Criterion corrected (AICc) of 108.0416. This value serves as a measure of the model's quality, balancing goodness of fit with simplicity.

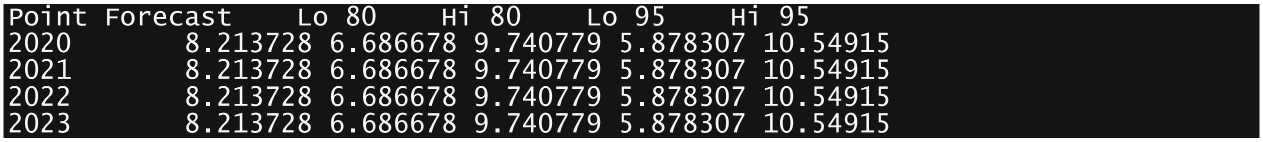
Subsequently, two more SES models were fitted with explicit smoothing factors to compare performance. The first of these, set alpha to 0.5, meaning that it equally weights the influence of the most recent observation and all previous observations. The AICc for this model was slightly higher at 111.7393, indicating a less efficient fit compared to the first model.

The third model used a higher smoothing factor of 0.75, giving more weight to the most recent observations. The summary output for this model wasn't fully detailed in the text but adjusting alpha to 0.75 typically aims to make the model more responsive to recent changes in the data pattern.

The formula for SES is:

$\hat{} yT+1|T=\alpha yT+\alpha(1-\alpha)yT-1+\alpha(1-\alpha)2yT-2+\cdots$,

The forecast are:



```
Point Forecast     Lo 80    Hi 80    Lo 95    Hi 95
2020       8.213728 6.686678 9.740779 5.878307 10.54915
2021       8.213728 6.686678 9.740779 5.878307 10.54915
2022       8.213728 6.686678 9.740779 5.878307 10.54915
2023       8.213728 6.686678 9.740779 5.878307 10.54915
```

Holt’s linear trend model

Two Holt’s linear trend models were applied to the logarithmically transformed EPEST_HIGH_KG data, which extends the simple exponential smoothing approach by explicitly incorporating a trend component in the model. The first model incorporated a damping parameter, making it a damped trend model, which adjusts the trend over time to become flatter, a useful feature when forecasting long-term where trends might not continue indefinitely. The AICc value for this damped model was 116.5141, indicating its fit and complexity balance.

The second model did not include a damping parameter and was also set to forecast 4 periods ahead. This model assumes that the trend will continue at a constant rate into the future. It yielded a lower AICc value of 112.4651 compared to the damped model, indicating a better fit to the data when not accounting for a reducing trend.

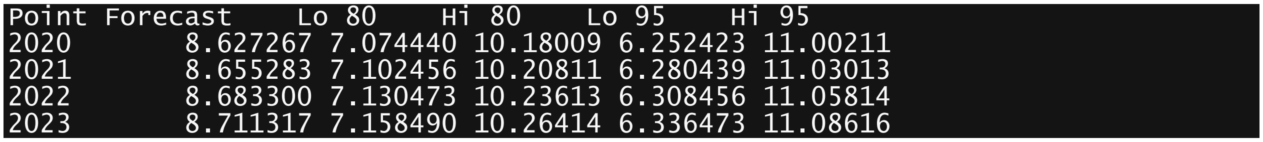
The formula for forecasting Holt’s linear trend model

Forecast equation: **$yt+h|t=\ell t+hbt$**

Level equation: **$\ell t=\alpha yt+(1-\alpha)(\ell t-1+bt-1)$**

Trend equation: **$bt=\beta*(\ell t-\ell t-1)+(1-\beta*)bt-1$,**

The forecast are:



```
Point Forecast     Lo 80    Hi 80    Lo 95    Hi 95
2020       8.627267 7.074440 10.18009 6.252423 11.00211
2021       8.655283 7.102456 10.20811 6.280439 11.03013
2022       8.683300 7.130473 10.23613 6.308456 11.05814
2023       8.711317 7.158490 10.26414 6.336473 11.08616
```

ETS (Exponential Smoothing State Space) models

For forecasting, four different configurations were tested on the logarithmically transformed `EPEST_HIGH_KG` data, aiming to identify the most effective model based on their AICc values.

The first model automatically selected the best fitting ETS model parameters without specific restrictions on the error, trend, or seasonality components. This model achieved the best performance with an AICc of 108.0415, indicating it provided the most statistically efficient fit among the tested models.

The second model was specified as an "ANN" model, which stands for Additive Error, No Trend, No Seasonality. This simpler model, focusing solely on the level component without incorporating trend or seasonal adjustments, yielded a nearly identical AICc of 108.0416, suggesting it performed almost as well as the best model.

The third model used a "MAN" configuration, which stands for Multiplicative Error, Additive Trend, No Seasonality. This model's AICc was 112.5402, indicating a less efficient fit compared to the first two models, likely due to the inclusion of an unnecessary trend component that didn't improve model performance according to the data characteristics.

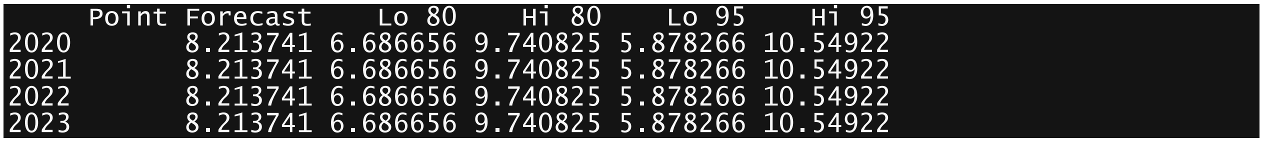
Finally, the fourth model specified as "MMN", incorporating Multiplicative Error, Multiplicative Trend, and No Seasonality, resulted in an AICc of 114.5862. This model, suggesting a more complex interaction of error and trend, performed the least effectively, which could be attributed to an overfitting issue or an inappropriate model structure for this data set.

Overall, the analysis shows that simpler ETS models, particularly those focusing primarily on the level without a trend component, were more effective for this dataset, as indicated by their lower AICc values. These results underscore the importance of model selection based on the data characteristics and the principle of parsimony, favoring simpler models when they fit the data well.

Formula for forecasting ETS model is:

et=yt−ℓt−1=yt−^yt|t−1 is the residual at time t**.**

The forecast are:



```
     Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
2020       8.213741 6.686656 9.740825 5.878266 10.54922
2021       8.213741 6.686656 9.740825 5.878266 10.54922
2022       8.213741 6.686656 9.740825 5.878266 10.54922
2023       8.213741 6.686656 9.740825 5.878266 10.54922
```

ARIMA model:

In a comprehensive analysis using ARIMA modeling on the logarithmically transformed `EPEST_HIGH_KG` data, several models were evaluated to determine the best fit based on the Akaike Information Criterion (AIC) values and other performance metrics such as RMSE and MAE.

The initial model specified as ARIMA (1,0,0), resulted in an AIC of 92.29, indicating a relatively simple model involving one autoregressive term. This model serves as a baseline for comparison with more complex specifications.

The model ARIMA (1,0,1), demonstrated the best overall performance with the lowest AIC of 91.4. It also performed better in terms of lower RMSE and MAE compared to other models, and its residuals were more characteristic of white noise, indicating a good fit to the data.

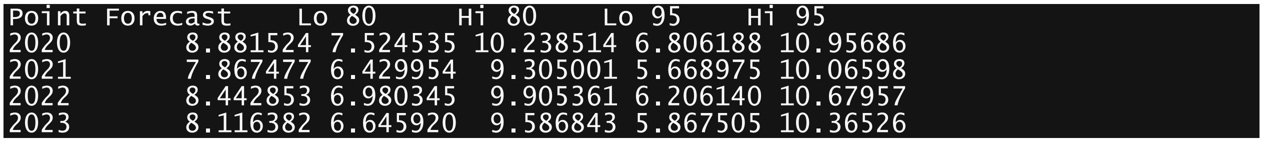
Lastly, ARIMA was generated by the `auto. Arima` function, which automatically selects the best model based on AIC, resulted in the lowest AIC of 91.2 and an AICc of 91.68, narrowly outperforming by a small margin in AIC, though both models show similar levels of performance.

In conclusion both exhibit robust performance, with excellent diagnostic metrics and achieving the lowest AIC, the choice between them may depend on specific needs for parsimony or slight differences in performance metrics. Both models represent efficient options for modeling and forecasting pesticide usage levels.

Formula for forecasting ARIMA model is:

$yt=c+\phi 1yt-1+\phi 2yt-2+\cdots+\phi pyt-p+\varepsilon t$,

The forecast are:



```
Point Forecast     Lo 80     Hi 80    Lo 95    Hi 95
2020       8.881524 7.524535 10.238514 6.806188 10.95686
2021       7.867477 6.429954  9.305001 5.668975 10.06598
2022       8.442853 6.980345  9.905361 6.206140 10.67957
2023       8.116382 6.645920  9.586843 5.867505 10.36526
```

Dynamic regression model:

In the analysis of the dynamic regression models applied to the logarithmically transformed `EPEST_HIGH_KG` data, using `EPEST_LOW_KG` as an exogenous regressor, a series of models were evaluated to determine their effectiveness based on the Akaike Information Criterion corrected (AICc). The models varied primarily in the complexity of their autoregressive (AR) and moving average (MA) terms.

The investigation began with simpler models and systematically increased in complexity to assess whether additional AR or MA terms would lead to improved model fit. The AICc values ranged significantly, indicating varying degrees of model efficiency across the different specifications. Lower AICc values typically suggest a better balance between model complexity and fit to the data.

The models tested showed a general trend were increasing the number of MA terms initially led to a decrease in the AICc, suggesting improved model fit. However, as more terms were added beyond a certain point, the improvements in AICc values began to diminish, and in some cases, the AICc increased, indicating potential overfitting.

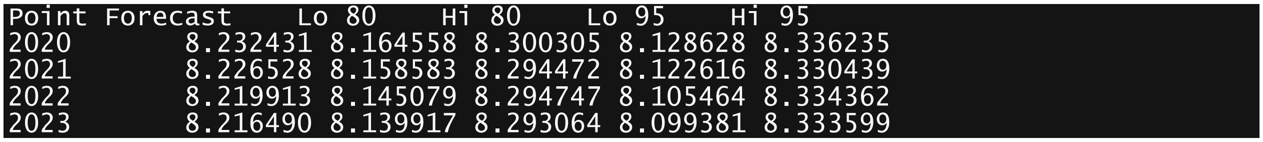
Further analysis of the residuals from these models revealed whether the assumptions of normally distributed and independent errors were met, which is crucial for the validity of the forecasts generated from these models. The best-fitting models typically exhibited residuals that appeared to be white noise, suggesting that the model was capturing most of the information in the data and leaving behind random fluctuations.

Overall, the dynamic regression approach allowed for a nuanced understanding of how `EPEST_LOW_KG` influenced `EPEST_HIGH_KG`, providing valuable insights into the relationship between different levels of pesticide usage. The exploration of various model configurations helped identify the optimal balance between model complexity and predictive accuracy, guiding effective decision-making for pesticide usage management and policy formulation.

The formula for forecasting dynamic regression are:

$yt=\beta0+\beta1x1,t+\cdots+\beta kxk,t+\eta t,(1-\phi1B)(1-B)\eta t=(1+\theta1B)\varepsilon t,$

The forecast are:

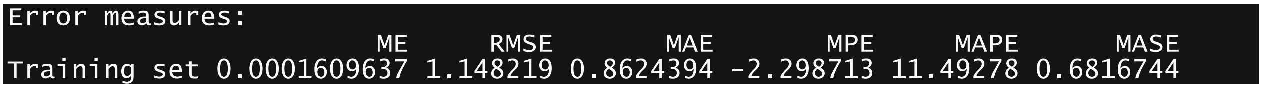


```
Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
2020       8.232431 8.164558 8.300305 8.128628 8.336235
2021       8.226528 8.158583 8.294472 8.122616 8.330439
2022       8.219913 8.145079 8.294747 8.105464 8.334362
2023       8.216490 8.139917 8.293064 8.099381 8.333599
```

5.RESULTS:

Simple Exponential Smoothing (SES)

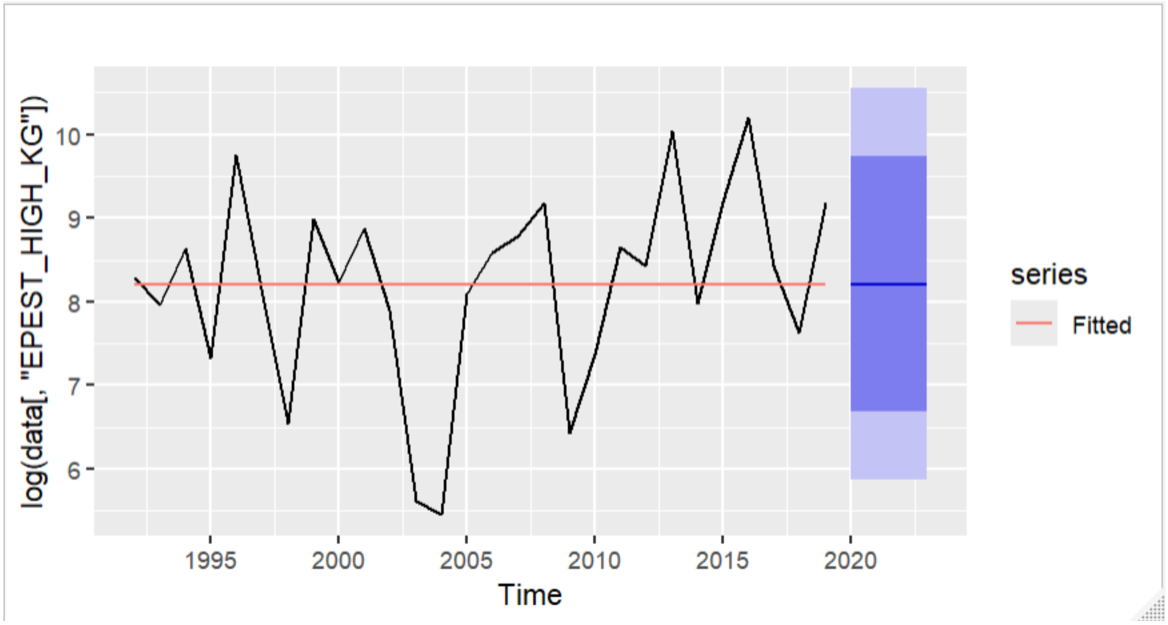
Forecasted graph for the next four years.



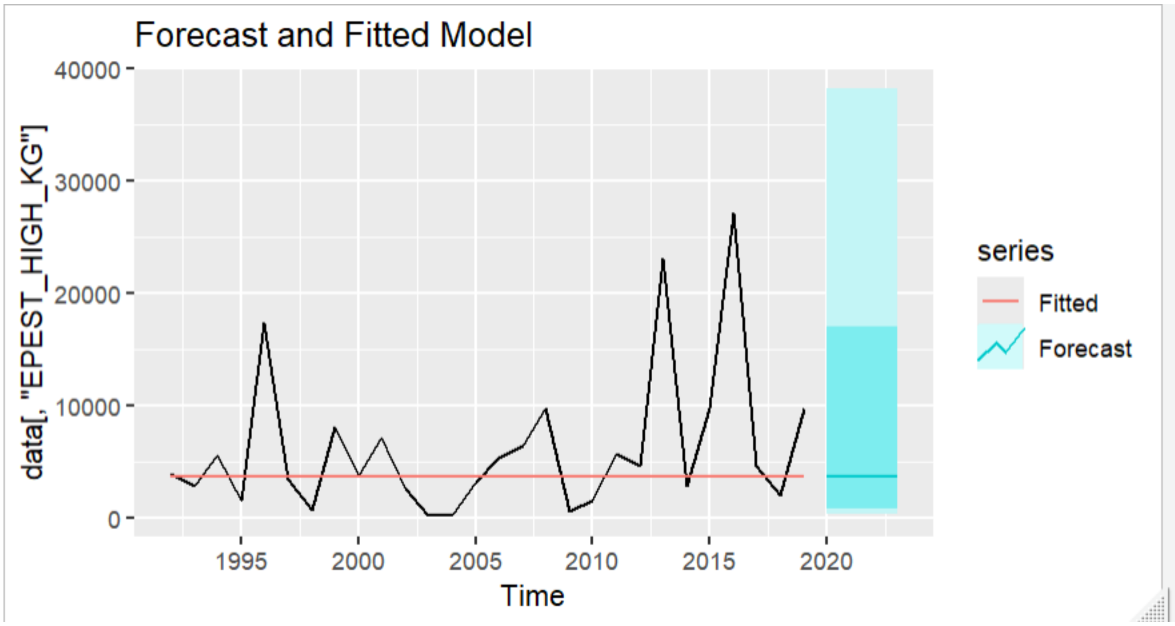
```
Error measures:
                       ME     RMSE       MAE       MPE     MAPE      MASE
Training set 0.0001609637 1.148219 0.8624394 -2.298713 11.49278 0.6816744
```

Forecast graph:

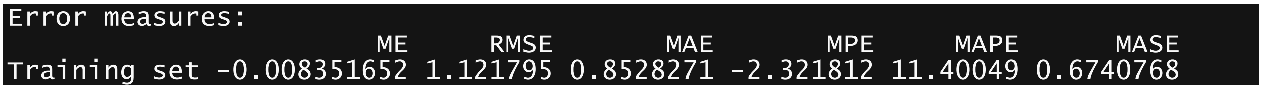
With log data (before back transformed)



With original data (after back transformed)



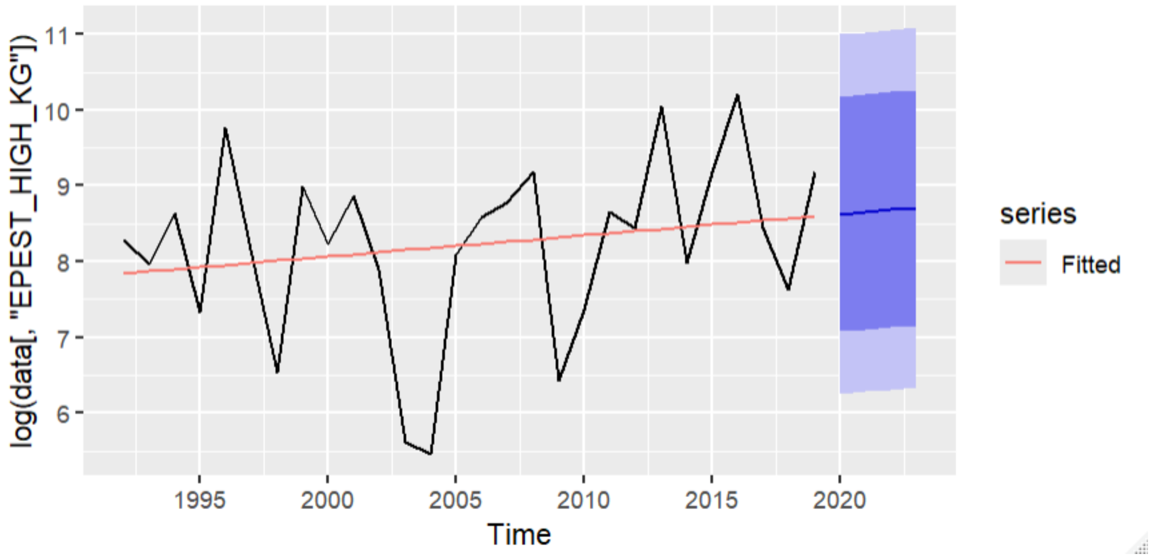
Holts model:



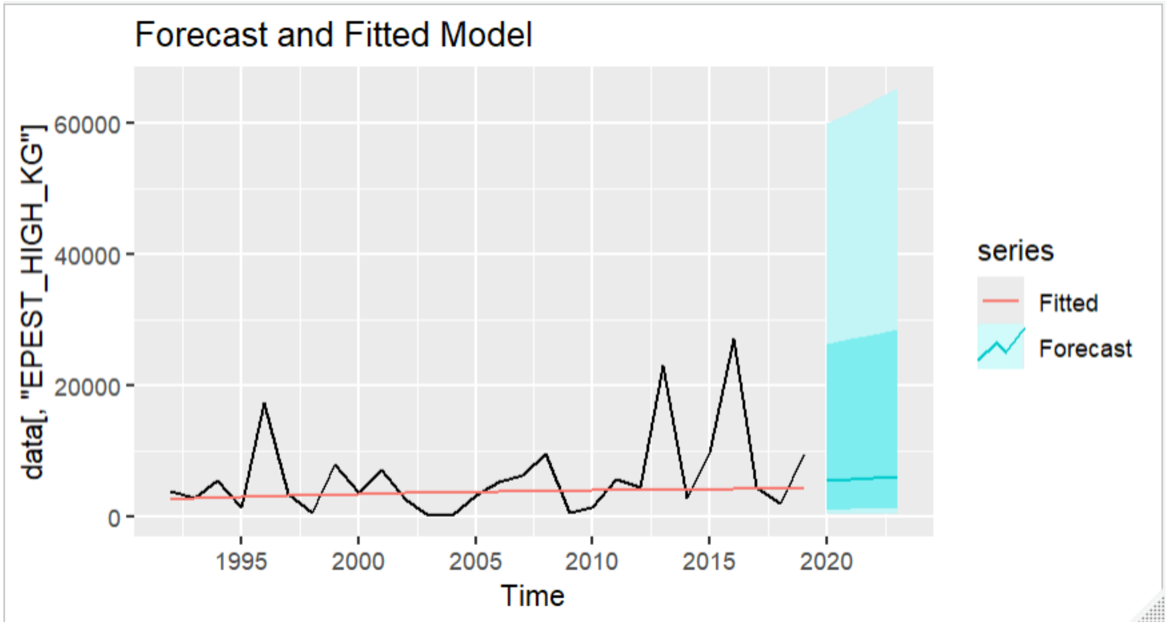
```
Error measures:
                       ME     RMSE       MAE       MPE     MAPE      MASE
Training set -0.008351652 1.121795 0.8528271 -2.321812 11.40049 0.6740768
```

Forecast graph:

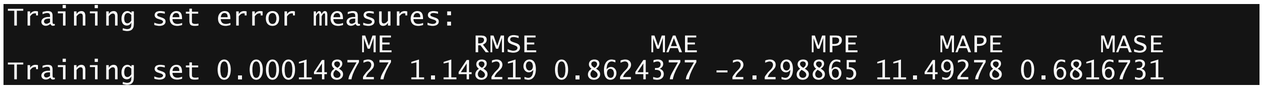
With log data(before back transformed):



With original data (after back transformed):

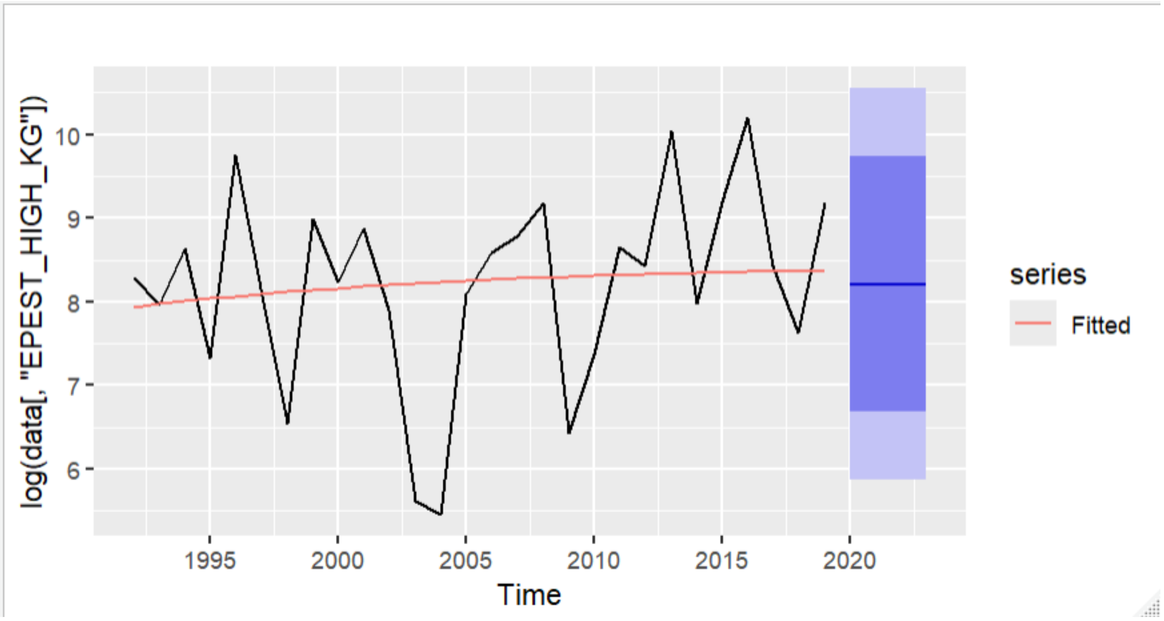


ETS (Exponential Smoothing State Space) models

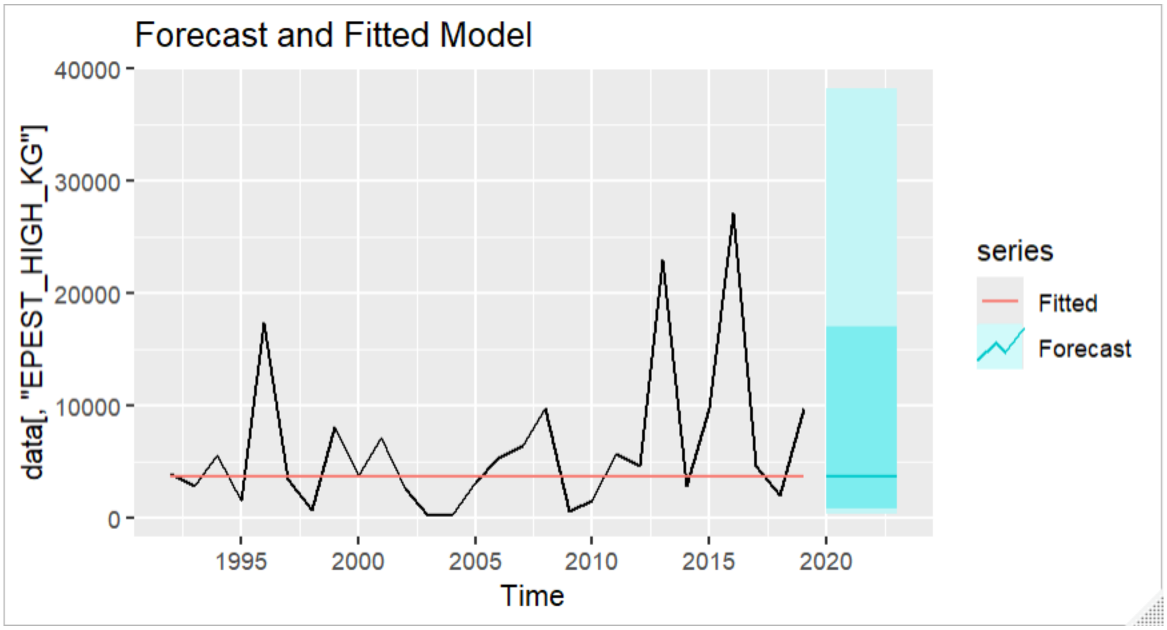


```
Training set error measures:
                     ME     RMSE       MAE       MPE     MAPE      MASE
Training set 0.000148727 1.148219 0.8624377 -2.298865 11.49278 0.6816731
```

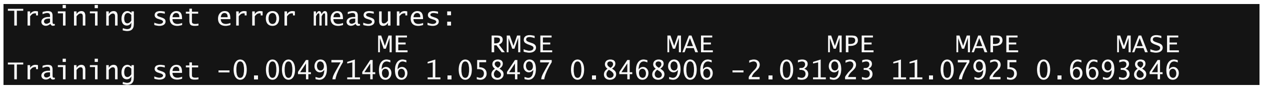
With log data (before back transformed):



With original data (after back transformed):

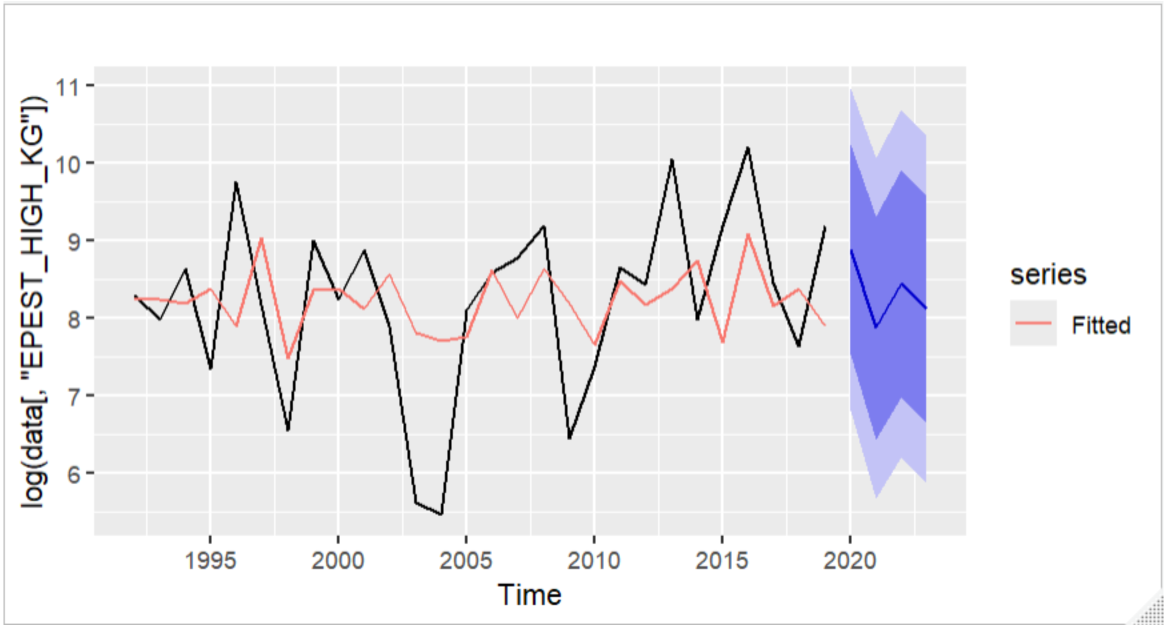


ARIMA model:

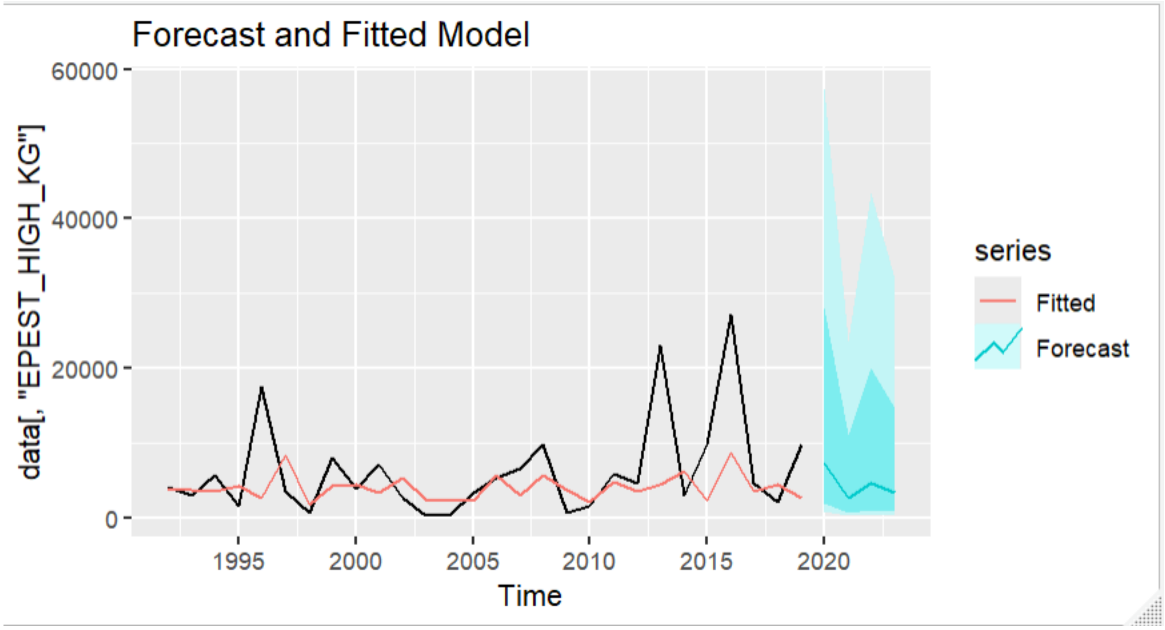


```
Training set error measures:
                       ME     RMSE       MAE       MPE     MAPE      MASE
Training set -0.004971466 1.058497 0.8468906 -2.031923 11.07925 0.6693846
```

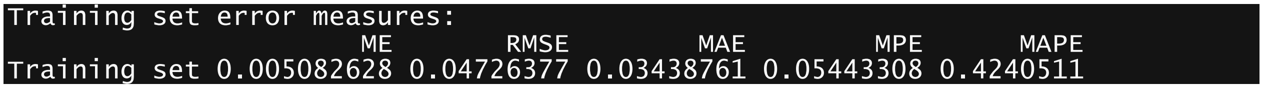
With log data (before back transformed):



With original data (after back transformed):

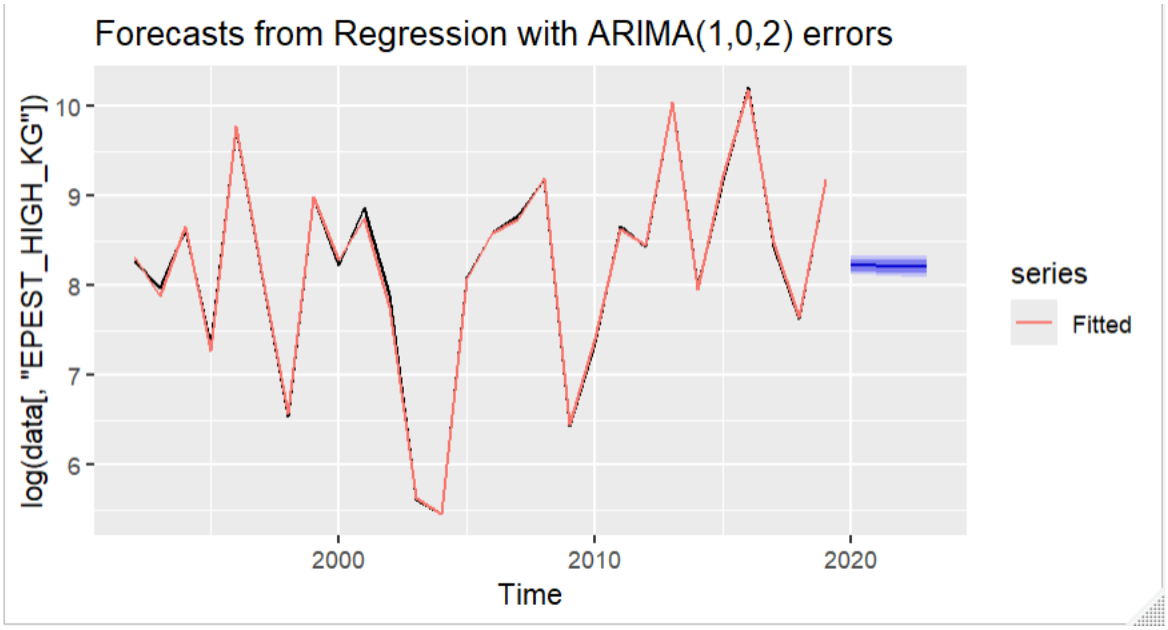


Dynamic regression model:

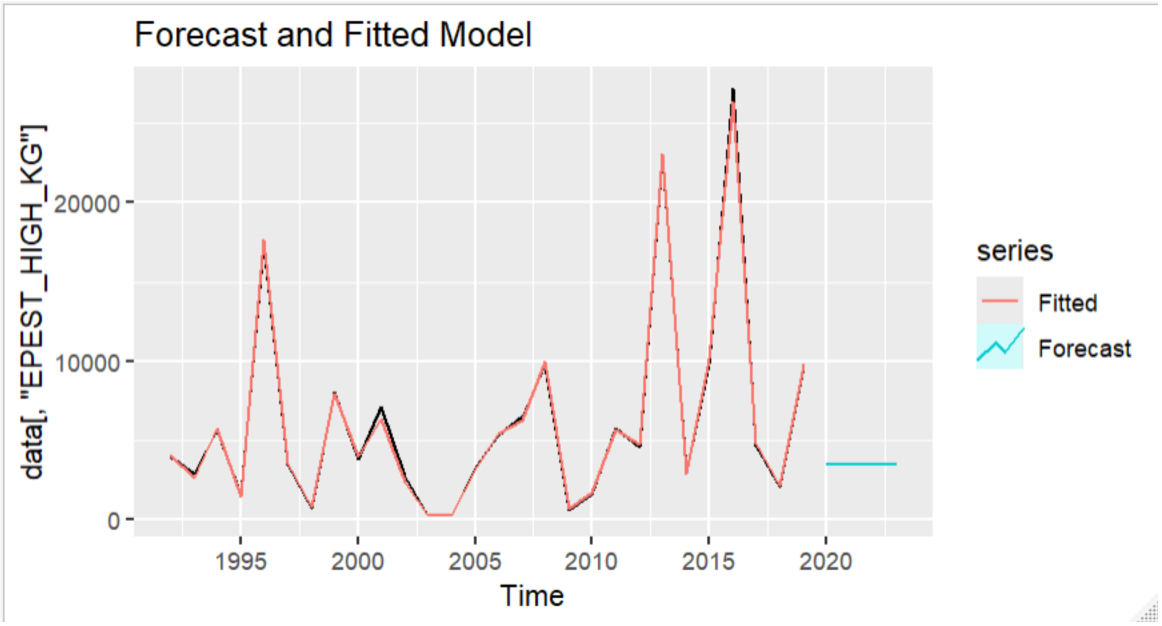


```
Training set error measures:
                     ME       RMSE        MAE        MPE      MAPE
Training set 0.005082628 0.04726377 0.03438761 0.05443308 0.4240511
```

With log data (before back transformed):



With original data (after back transformed):



CONCLUSION:

We utilized historical data from the dataset to forecast pesticide usage for the next four years. Employing five different strategies, we found that ARIMA forecasting demonstrated superior performance compared to the alternatives.

After conducting back transformed accuracy testing on both the test and training sets mentioned above, it was observed that the RMSE values for ARIMA were significantly lower compared to other models. Consequently, ARIMA emerged as the top-performing model for predicting pesticide usage across the US.

MODEL	Accuracy test (RMSE)
SES	12393.688659
HOLT	12169.896890
ETS	12394.036822
ARIMA	10530.69687
Dynamic regression	12416.01

Notably, the presence of two spikes towards the end of the original data, which was utilized as training data, resulted in significant errors across all models. However, by disregarding these spikes, the algorithm shows promise in effectively predicting pesticide usage in the future.

CITATIONS:

https://web-p-ebscohost-com.ezproxy.uta.edu/chc/pdfviewer/pdfviewer?vid=0&sid=3dcc452f-921c-4736-a302-124db743df0a%40redis

https://web-p-ebscohost-com.ezproxy.uta.edu/ehost/pdfviewer/pdfviewer?vid=0&sid=4c978aaf-c2a0-4b98-a89a-d3066c24b361%40redis

https://www.niehs.nih.gov/health/topics/agents/pesticides#:~:text=Pesticides%20kill%2C%20repel%2C%20or%20control,are%20commonly%20used%20on%20lawns.

https://en.wikipedia.org/wiki/Pesticide

https://web-p-ebscohost-com.ezproxy.uta.edu/bsi/pdfviewer/pdfviewer?vid=4&sid=1656cde0-bdc9-4b1b-8f7b-6f9495b411ed%40redis

https://web-p-ebscohost-com.ezproxy.uta.edu/ehost/pdfviewer/pdfviewer?vid=0&sid=f0ac352a-c4e5-4f8d-9d76-7e5f5e87dcd5%40redis

https://www.sciencedirect.com/science/article/abs/pii/S0048969717302784

https://www.sciencedirect.com/science/article/abs/pii/S0301479712004689

https://www.mdpi.com/1660-4601/8/5/1402