Modeling Year-Over-Year Changes in Global Lithium Production

Klarissa Castillo, Paulo Cataquis, Diego Coroando, Cameron Htut, Syrus Tolentino

July 23, 2025

Introduction

There are a lot of resources on Earth for energy applications and storage, but not many can also be applied to medicine or textiles. One that stands prevalent is lithium, a non-renewable metal resource. With the majority of lithium products being batteries for cars, electronics, and energy storage; as well as medication for bipolar disorders, lithium serves a multitude of applications. In the past couple decades, due to its various applications, the supply, demand, and consumption of lithium has soared. Due to its finite nature, we ask ourselves, how has global production for this commodity changed annually and what can we infer about future lithium production.

To do this, we analyzed global lithium production data from 2019 to 2024, collected from the U.S. Geological Survey (USGS). The data includes 540 observations from four variables:

- Entity – The country of origin of Lithium
- Code – Country Code
- Year - The year recorded per observation
- Lithium Production (KT) - Numerical value of lithium extracted (in KiloTonnes)

Input Variable: `Year` (centered as `year_c`)

Response Variable: `production` (global production in KT)

Primary Question:

How has total world lithium production changed year over year?

Methods and Results

Linear Regression Model (Paulo, Klarissa)

To understand how global lithium production has changed year over year, we applied a Linear Regression Model. Our response variable is Production, representing total worldwide lithium output in kilotons. The predictor we used was Year_c, a centered version of the calendar year

(Year - 2019). This transformation improves interpretability by aligning the intercept with the base year 2019. We selected linear regression because it is one of the most interpretable and reliable models for analyzing simple time-based trends. In our case, we were not only interested in predicting values but also in estimating the average change in lithium production per year. Some advantages are that they are simple and fast to compute, produce a direct slope estimate for a year over year change, and easy to interpret in real world terms. Disadvantages are that they assume a constant linear relationship between the predictor and response, it can underfit when the data includes sudden changes or nonlinear jumps and are sensitive to outliers like our drop in 2024

The linear regression formula would be:

$$\text{Production} = \beta_0 + \beta_1 \cdot \text{Year\_c}$$

We considered only one predictor: Year_c. Additional predictors like price, country reserves, or imports were excluded because the dataset has only six observations, and adding more variables would drastically increase the risk of overfitting. Our goal was to assess the trend in global production, not build a multivariate explanatory model. Including more predictors would have required testing their p-values for significance, and with such a small sample size, those values would likely be unstable and unreliable. Thus, linear regression with a single predictor was the most appropriate approach for our research question and data limitations.

In each iteration, we randomly selected 80% of the data to serve as the training set and used the remaining 20% as the test set. The model was fit on the training data using the formula [Equation] which predicts global lithium production based solely on the centered year variable. After training the model, we generated predictions for the test set and calculated the root mean squared error (RMSE) between the predicted and actual production values. This process was repeated across 10 different random splits, and we recorded the RMSE from each run. We used a fixed random seed (set.seed(123)) to ensure reproducibility of the results. This approach provided a robust estimate of how well the model generalizes to unseen data while mitigating the randomness associated with a single split. The average test RMSE across the 10 iterations was approximately 383,598.8 kilotons, suggesting that the model's typical prediction error is around 383,000 kilotons. Given the very small size of the dataset (n = 6) and the presence of an outlier in the 2024 estimate, this result is acceptable and demonstrates the model's capacity to capture the general trend in lithium production over time.

Tree Based Methods (Cameron, Diego, Syrus)

To explore alternative modeling approaches beyond linear regression, we first considered decision trees. A decision tree recursively splits the data into regions based on threshold values of the predictor. This flexibility allowed the decision tree to capture potential nonlinear patterns that a linear model might miss, even with only one predictor.

However, a major limitation of decision trees is their instability. Small changes in the data can lead to large changes in the tree structure. This was especially problematic in our dataset, which contained few observations and includes an anomaly in 2024 due to missing U.S. data. This kind of sensitivity often results in overfitting and poor generalization.

The following plot shows predicted values vs. actual production, highlighting the model's step-like predictions that fail to capture the smooth upward trend in production. Additionally, we visualized the cross-validation deviance and resulting pruned tree to demonstrate its instability.

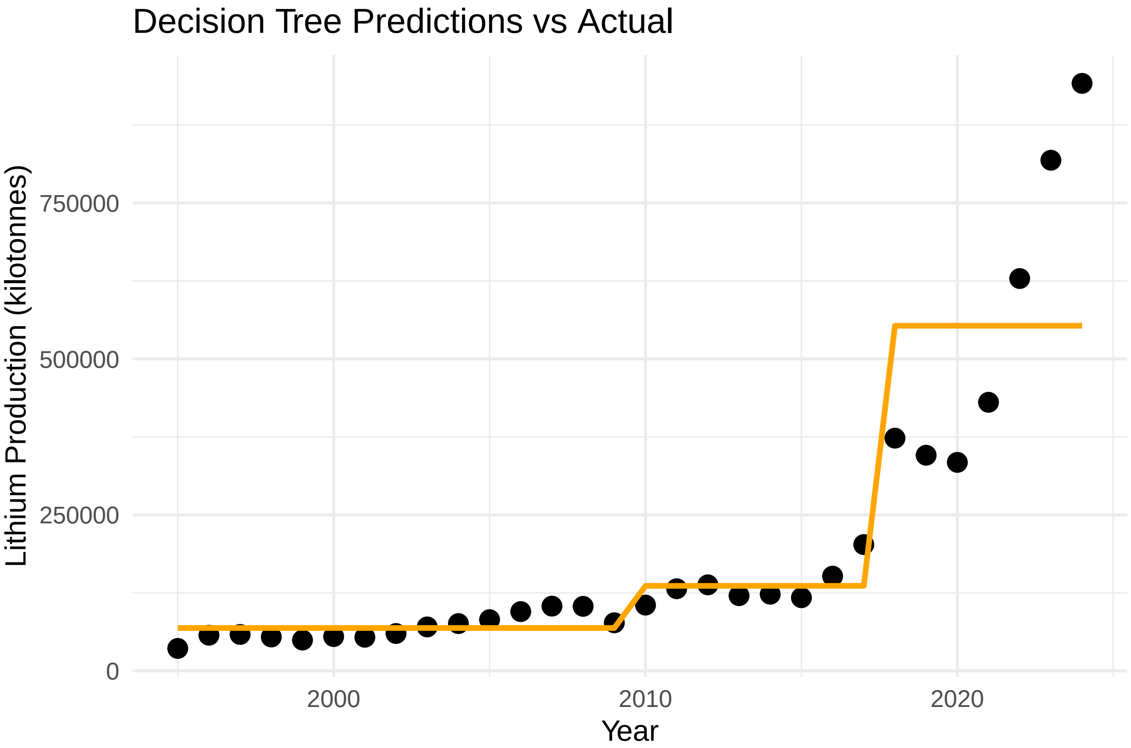


Figure 1: Decision Tree Predictions vs Actual Production. The model fails to capture smooth trends and introduces abrupt shifts due to its limited depth and sensitivity to splits.

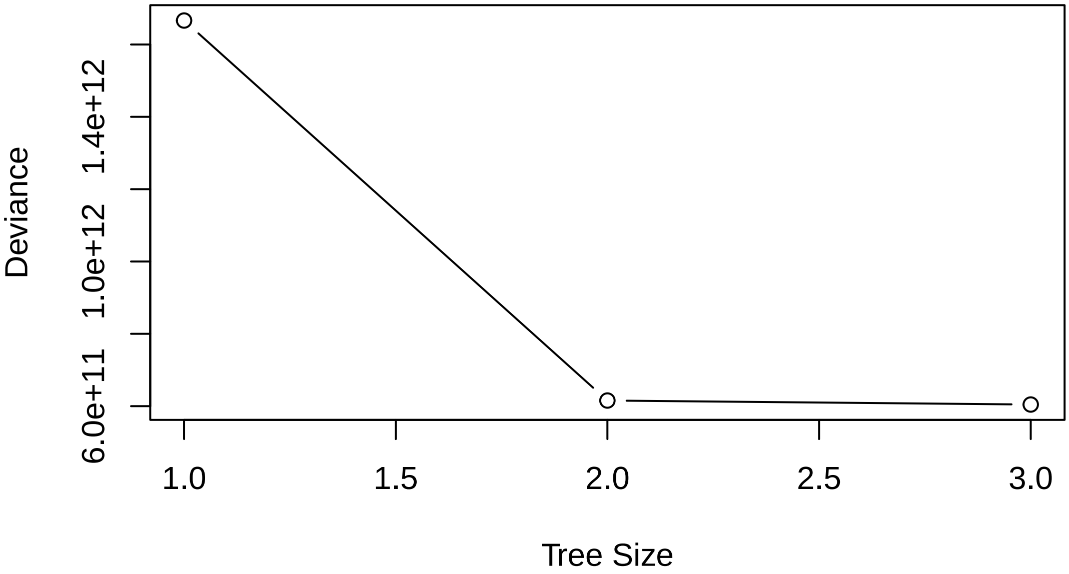


Figure 2: Cross-validation deviance for various tree sizes. Smaller trees may generalize better, but deviance remains high, indicating poor fit.

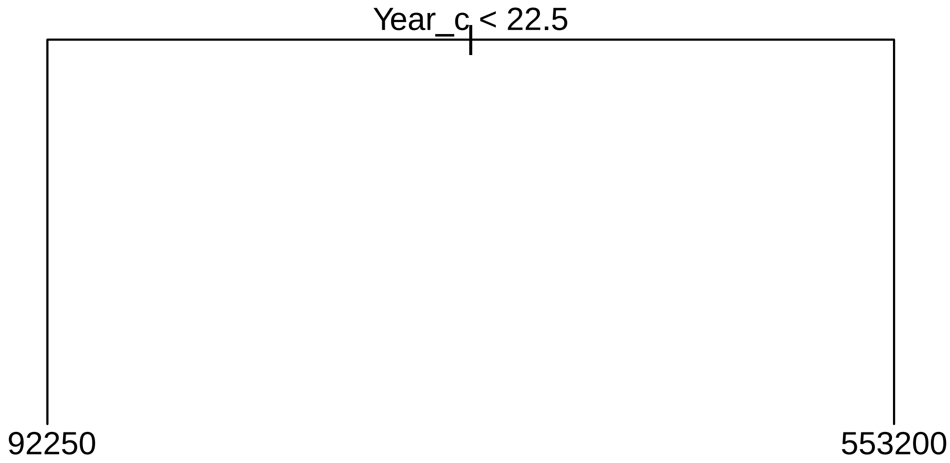


Figure 3: Pruned decision tree (size = 2). Despite pruning, the model captures only coarse patterns and misses gradual production trends.

To address these limitations, we used a **Random Forest**, an ensemble method that fits multiple decision trees on different bootstrapped subsets of the data and averages their predictions. This significantly improves model stability while retaining the flexibility to capture nonlinear patterns.

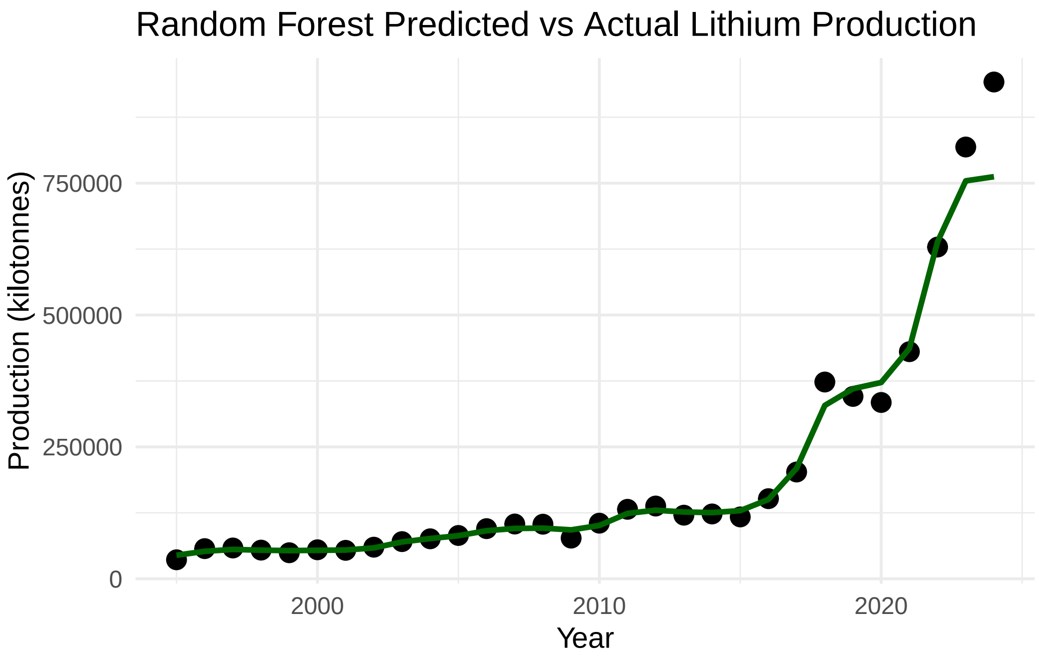
We trained a random forest model using 5-fold cross-validation with 500 trees (`ntree = 500`). The model outperformed all others in terms of RMSE, suggesting it struck the best balance between flexibility and generalization.

```
train_ctrl <- trainControl(method = "cv", number = 5)

rf_model <- train(Production ~ Year_c, data = df_global, method = "rf",
                  trControl = train_ctrl, ntree = 500)
rf_model
```

```
Random Forest

30 samples
 1 predictor

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 23, 24, 25, 24, 24
Resampling results:

  RMSE      Rsquared   MAE
  61073.85  0.9822375  35701.03

Tuning parameter 'mtry' was held constant at a value of 2
```

```
df_global$pred_rf <- predict(rf_model, df_global)
```

```
ggplot(df_global, aes(x = Year)) +
  geom_point(aes(y = Production), size = 3) +
  geom_line(aes(y = pred_rf), color = "darkgreen", size = 1) +
  labs(
    title = "Random Forest Predicted vs Actual Lithium Production",
    y = "Production (kilotonnes)"
  ) +
  theme_minimal()
```



Results

Results

Conclusion

Conclusion