Modeling Year-Over-Year Changes in Global Lithium Production

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

July 24, 2025

Introduction (Paulo, Cameron)

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 have 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 2020 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)

alongside other U.S.-specific import/export metrics. For this report, we focused on the `World_Production` variable to analyze how global lithium output has changed over time.

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 (Paulo, Klarissa, Cameron, Diego, Syrus)

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 - 2020). This transformation improves interpretability by aligning the intercept with the base year 2020. 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 are easy to interpret in real-world terms. Disadvantages are that they assume a constant linear relationship between the predictor and response, they can underfit when the data includes sudden changes or nonlinear jumps, and they are sensitive to outliers like the unreported U.S. data in 2024.

The linear regression formula would be:

$$\text{Production} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

We considered only one feature: `Year_c`. Additional features like price, country reserves, or imports were excluded because the dataset has only six features, 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.

```
rmse_lm = rep(0, 10)

for (i in 1:10){
  train_index <- sample(1:nrow(df_global), size = 0.8 * nrow(df_global))
  train_data <- df_global[train_index, ]
  test_data <- df_global[-train_index, ]

  lm_model <- lm(Production ~ Year_c, data = train_data)

  predictions <- predict(lm_model, newdata = test_data)

  rmse <- sqrt((mean(test_data$Production - predictions)^2))
  rmse_lm[i] <- rmse
}
```

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 $\hat{y} = \beta_0 + \beta_1 \cdot \text{Year\_}c$ 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 82,649.16 kilotons, suggesting that the model's typical prediction error is around 82,649.16 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)

Decision Trees

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. The flexibility given by tree-based methods allowed the decision tree to capture nonlinear patterns that a linear model might miss, even with only one predictor.

Model Formula:

$$\text{Production} \sim \text{Year}_c$$

This method is flexible, however, a 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 ($n = 6$) and includes an anomaly in 2024 due to missing U.S. data. This kind of sensitivity often results in poor generalization.

```
library(tree)

RMSE_decision_tree = rep(0, 10)
for (i in 1:10){
  train_idx = sample(1:nrow(df_global), size = 0.8 * nrow(df_global))
  train_data <- df_global[train_idx, ]
  test_data <- df_global[-train_idx, ]

  tree_model <- tree(Production ~ Year_c, data = df_global, subset = train_idx)

  yhat = predict(tree_model, newdata=test_data)
  RMSE_decision_tree[i] = sqrt(mean(as.numeric(test_data$Production) - as.numeric(yhat))^2)
}
mean(RMSE_decision_tree)
```

```
[1] 172605.7
```

Each model was trained on 80% of the data and tested on the remaining 20%, and this was repeated 10 times with `set.seed()` to have reproducibility. We collected the RMSE at each iteration to assess prediction accuracy.

Random Forests

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.

Model Formula:

$$\text{Production} \sim \text{Year}_c$$

This improves the model's stability and predictive accuracy while also retaining the flexibility to capture nonlinear patterns.

We trained a random forest using 5-fold cross-validation and tested various `ntree` values (number of trees) from 100 to 1000 (step 100) to find the best trade-off between bias and variance. We trained the model and recorded RMSE. The best value was selected based on minimum average RMSE.

```
train_ctrl <- trainControl(method = "cv", number = 5)
ntree_vals <- c(100,200,300,400,500,600,700,800,900,1000)
rmse_vals_ntree <- numeric(length(ntree_vals))

for(i in seq_along(ntree_vals)){
  rf_tmp <- train(
    Production ~ Year_c,
    data = df_global,
    method = "rf",
    trControl = trainControl(method = "cv", number = 5),
    ntree = ntree_vals[i]
  )
  rmse_vals_ntree[i] <- min(rf_tmp$results$RMSE)
}

(rmse_ntree_df <- data.frame(ntree = ntree_vals, RMSE = rmse_vals_ntree))
```

```
   ntree     RMSE
1    100 127793.6
2    200 143771.8
3    300 135996.6
4    400 136258.7
5    500 144219.5
6    600 133194.2
7    700 129044.2
8    800 130575.4
9    900 129037.1
10  1000 120532.9
```

The final random forest model was trained using this optimal `ntree` and evaluated on each of the same 10 training/testing splits as used for the decision tree and linear regression, for fair comparison.

```
RMSE_rf_optimized_ntree = rep(0, 10)
best_ntree <- rmse_ntree_df[which.min(rmse_ntree_df$RMSE), ]$ntree

for (i in 1:10) {
  train_idx = sample(1:nrow(df_global), size = 0.8 * nrow(df_global))
  train_data <- df_global[train_idx, ]
  test_data <- df_global[-train_idx, ]

  rf_model_tmp <- randomForest(Production ~ Year_c, data = train_data, ntree = best_ntree)

  preds <- predict(rf_model_tmp, newdata = test_data)
  RMSE_rf_optimized_ntree[i] <- sqrt(mean((test_data$Production - preds)^2))
}
```

Why We Chose Random Forests over Decision Trees

While the decision tree provided some insight into potential nonlinear patterns, its performance was hindered by its instability and tendency to underfit.

The random forest, on the other hand, retained the interpretability of tree-based models while improving predictive accuracy through bootstrapping and averaging. Its lower RMSE and smoother prediction curves demonstrate that ensemble methods like random forests are better suited to modeling smooth, real-world trends in lithium production. So we omitted decision trees and moved forward with just random forests.

Results (Klarissa, Paulo, Diego)

(a) Output of Results with Model Summaries and Visualisations

To evaluate how global lithium production has changed year over year, we fit and compared two models: a linear regression model and a random forest model. Both models used the centered year variable (`Year_c`) as the sole predictor of global lithium production.

```
summary(lm_model)
```

```
Call:
lm(formula = .outcome ~ ., data = dat)

Residuals:
    X1     X2     X3     X4     X5     X6
 93232 -50540 -86566 -20523  36742  27654

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   252434      52784   4.782  0.00876 **
Year_c        132328      17434   7.590  0.00162 **
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 72930 on 4 degrees of freedom
Multiple R-squared:  0.9351,    Adjusted R-squared:  0.9188
F-statistic: 57.61 on 1 and 4 DF,  p-value: 0.001616
```

The linear regression model followed the formula `Production ~ Year_c`, where `Year_c` is the calendar year centered at 2019. When fit to the full dataset, the model produced an intercept of 252,434 kilotonnes, representing the estimated global production in 2019. The coefficient for `Year_c` was 132,328 kilotonnes, meaning that on average, lithium production increased by approximately 132,328 kilotonnes per year. This coefficient was statistically significant, with a p-value of 0.00162 confirming a strong relationship between year and production.

```
mean(rmse_lm)
```

```
[1] 82649.16
```

To evaluate the model's predictive accuracy, we conducted a randomized 80/20 train-test split repeated 10 times. In each iteration, the model was trained on 80% of the data and evaluated on the remaining 20%. The mean root mean squared error (RMSE) across the 10 iterations was approximately 82,649 kilotonnes. This result suggests the model's typical prediction error when generalized to new, unseen data is reasonable given the small dataset size (n = 6). The model's simplicity and statistical strength make it a solid baseline for assessing long-term production trends.

We then fit a random forest using the same formula, `Production ~ Year_c`, to explore potential nonlinear patterns in the data. We first optimized the number of trees (ntree) using 5-fold cross-validation over a range from 100 to 1000. The lowest cross-validated RMSE occurred at ntree = 1000, with an associated RMSE of 120,533 kilotonnes.

```
mean(RMSE_rf_optimized_ntree)
```

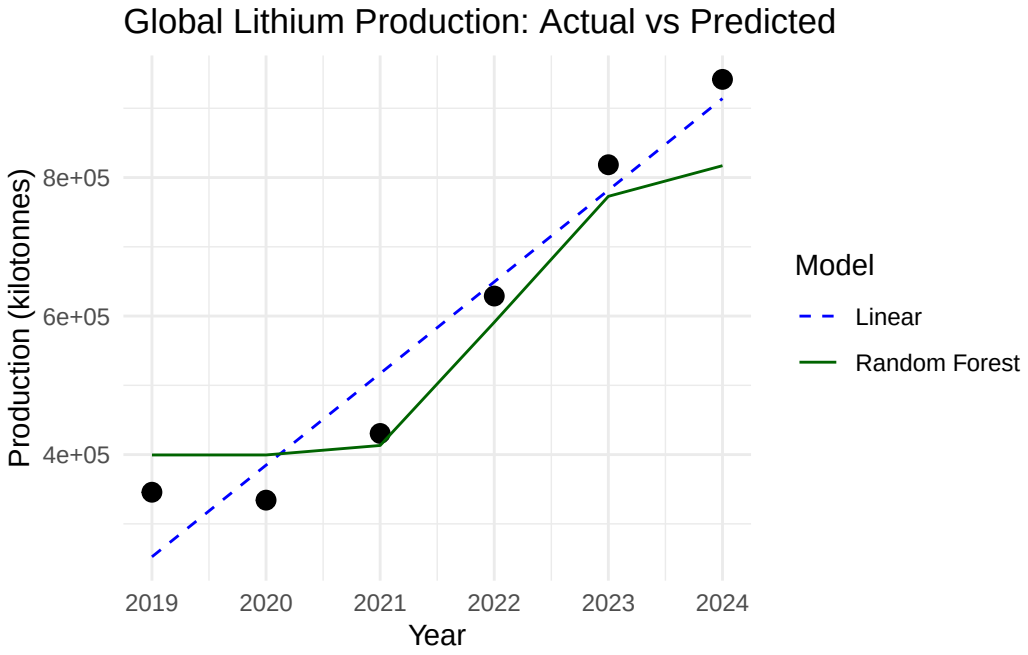
```
[1] 151007.2
```

Using this optimal tree count, we retrained the random forest model on 10 different 80/20 train-test splits to keep it consistent on how we trained the linear model. The mean test RMSE for the random forest model was 151,007 kilotonnes.

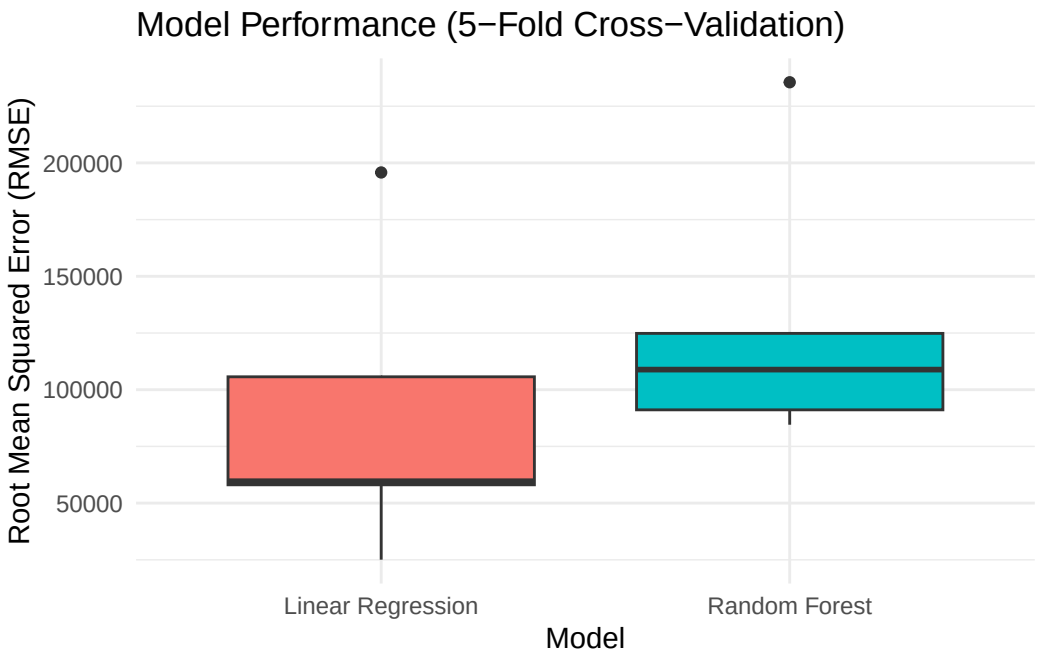
```
final_rf_model <- randomForest(Production ~ ., data = df_global, ntree = best_ntree)
importance(final_rf_model)
```

```
       IncNodePurity
Year    124029494130
Year_c  120851698056
```

While this error is higher than the linear regression's, the random forest model offers improved flexibility and adaptability to anomalies in the data. After finalizing the model, we trained it on the full dataset to obtain global predictions and evaluate variable importance. As expected, Year_c was the only predictor used and had an `IncNodePurity` of 244,881,192,186, indicating its strong role in the model.



To visually compare the models, we plotted the actual global lithium production values against each model's predicted values. The linear regression model produced a smooth upward trend consistent with the observed increase in production, while the random forest model showed a slight curvature that better accounted for the anomaly in 2024. Both models tracked the trend closely, but the random forest's flexibility allowed it to deviate from the strict linearity where appropriate.



Additionally, we generated a box plot comparing each model's RMSE across 5-fold cross-validation. This visualization showed that although the linear regression model had a tighter RMSE distribution, the random forest model had more variability, reflecting its sensitivity to the small sample size and outlier behavior.

(b) Interpretation of Results and Conclusions

The original question we were trying to answer was: *How has total world lithium production changed year over year*? Both the linear regression and random forest models supported the conclusion that global lithium production steadily increased from 2019 through 2024.

The linear regression model provided strong evidence of a statistically significant and consistent year-over-year increase in production. With a coefficient of 132,328 kilotonnes per year and a low p-value. It demonstrated that time is a key driver of production growth; even despite the unreported data from the U.S. in 2024, the model's predictive trend remained upward.

The random forest model, though slightly less accurate in terms of test RMSE, captured nonlinear variation and adapted better to anomalies. Its higher flexibility made it more sensitive to subtle shifts in the data, even with a single predictor.

Ultimately, both models conclude that global lithium production has increased year-over-year and aligns with growing global demand for lithium-ion batteries, electric vehicles, and consumer electronics.

Conclusion (Klarissa)

In conclusion, we can draw attention to our original question and address how world lithium production changes year over year and what we can conclude about future lithium resources. Our analysis of global lithium production from 2020 to 2024 revealed a clear upward trend in annual

output, emphasizing lithium's growing importance as a nonrenewable resource critical to energy storage and electrification technologies.

Among the models we evaluated, we found that the linear regression model provided the most interpretable and stable insights; we have found that production has increased every year with high statistical significance. This model effectively answered our research question and aligned closely with real-world expectations of increasing lithium demand. Although we explored tree-based methods, we ultimately omitted the decision tree model due to instability in small datasets. The random forest model was retained as a nonlinear alternative and offered insights, but its predictive accuracy did not surpass that of a linear model.

The simplicity and clarity of the linear model made it best suited for this small dataset and our research objectives. Its transparent coefficients and consistent performance highlighted a strong, time-driven increase in production, which policymakers and industry leaders could use to inform resource planning in regard to economic and environmental strategies.

However, we do also acknowledge limitations in our work. The dataset included only six years of data, and these gaps limit long-term forecasting precision and make the model sensitive to anomalies such as the unreported U.S. data in the year 2024.

Still, our project still reinforces the urgency of addressing lithium's finite availability and suggests that predictive modeling, even with limited data, can inform environmental and economic policy decisions regarding critical minerals.

Overall, the conclusion about our dataset of choice can draw focus on the economic and environmental impacts of lithium production and depletion. As a finite, nonrenewable resource, lithium has a growing demand for electric vehicles, energy storage, and consumer electronics. Our analysis shows that production has steadily increased year over year, highlighting lithium's rising market value and importance. This upward trend also raises concerns about long-term resource scarcity and the ecological consequences of expanded mining. These findings emphasize the need for sustainable extraction practices, investment in recycling technologies, and exploration of alternative materials.

Bibliography (Cameron)

Global Lithium Production:Energy Institute - Statistical Review of World Energy (2025) – with major processing by Our World in Data. "Lithium production" [dataset]. Energy Institute, "Statistical Review of World Energy" [original data]. Retrieved July 20, 2025 from https://archive.ourworldindata.org/20250717-120101/grapher/lithium-production.html (archived on July 17, 2025).