W271 Lab 2: CO2 Present

Contents

0.1 (3 points) Task 0b: Introduction

In our 1997 report, we developed the linear models and the ARIMA model to forecast the CO_2 level through 2020 and beyond. The CO_2 level increased from 360 ppmv in December 1997 to 420 ppmv in February 2023. The 420 ppmv level was at the higher end of our forecast range (385-424 ppmv) with 95% confidence. In this follow-up study, we will compare our 1997 forecast to the actuals, and make further enhancements to the model performance.

The original data set was a monthly series and the new data set has both monthly and weekly frequency. Also, due to the eruption of the Mauna Loa Volcano, measurements from Mauna Loa Observatory were suspended as of 11/29/2022. Observations starting in December 2022 are from a site at the Maunakea Observatories, ~21 miles north of the original site. We believe the site change did not pose a significant impact on the data generation process.

0.2 (3 points) Task 1b: Create a modern data pipeline for Mona Loa CO2 data.

We sourced the weekly and monthly data set from the United States' National Oceanic and Atmospheric Administration data page [here]. We noted 18 observations with -999.99 value (missing values) in the weekly dataset. We replaced the 18 invalid weekly values with their corresponding month's value in the monthly data set. Then we generated two clean time series data sets from May 1974 to March 2023: co2_present (weekly frequency) and co2_present_month (monthly frequency).

Similar to our 1997 observations, there is an increasing CO_2 trend and seasonal variability in the data set. See charts in Figure 1. However, the trend line seems to steepen in after mid-2000s. The annual growth rates are mostly in the range of 0 to 1%, with a modest upward trend (also observed in 1997). We noticed that the long run average of the annual growth rates now is 0.495%, higher than the prior long run average of annual growth rate of 0.37% in the 1997 data set. This further confirms that CO_2 levels increased faster in the recent years. The CO_2 Decomposition graph continued to show an upward trend, strong seasonality, and the irregular effect.

We also noted that the peak CO_2 month is April (instead of May in the 1997 report) and the lowest month is September (instead of October). The seasonal difference between the peak and trough month is (6.185 ppmv), similar to the 1997 observation.

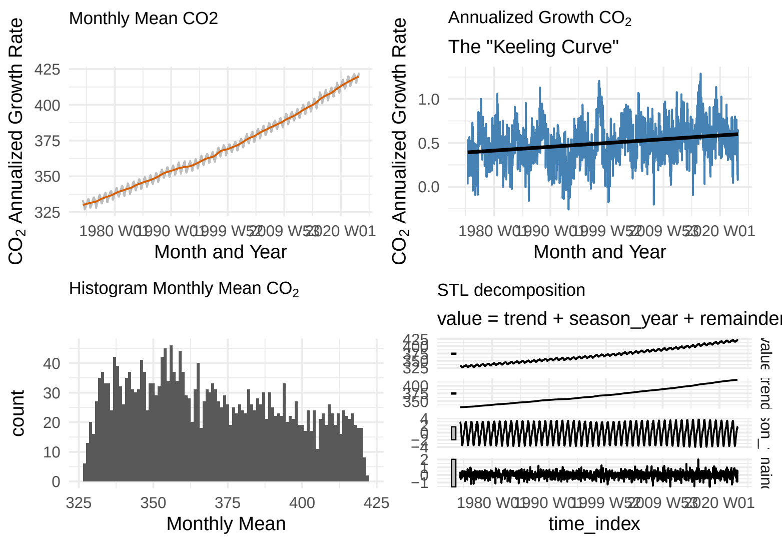

Figure 1: Atmospheric CO2 Level Time Series Overview

0.3 (1 point) Task 2b: Compare linear model forecasts against realized CO2

In our 1997 report, we developed the two linear models: a linear model using time as the single variable, and a polynomial model using the third polynomial degree of trend and seasonality. We used these models to forecast the CO_2 level through 2020 and beyond. In comparison of the forecast results to the actuals through the present (February 2023), we noted that both models under-forecasted the CO_2 levels. In a relative term, the linear model outperformed the polynomial model in the long run, which indicated the overfitting issues of the third degree polynomial model. See the comparison in Figure **??**.

The actual level CO_2 was 420 ppmv in February 2023. The linear model projected a straight trend line (without any seasonal effect) and forecasted CO_2 to reach 395 ppmv in February 2023. The polynomial model captured the seasonal effect and projected increasing levels until the mid-2021 and then flattened and shifted downward. The polynomial model forecasted 384 ppmv in February 2023.

0.4 (1 point) Task 3b: Compare ARIMA models forecasts against realized CO2

In our 1997 report, we also fitted an ARIMA model (ARIMA(0,1,1)(1,1,2)[12]). This model captured the seasonal effect and projected well until 2005 but under-forecasted after 2005. The ARIMA model forecasted 405 ppmv for 2/2023 vs. the actual level of 420 ppmv (figure). Relatively, the ARIMA model performs much better than the two linear models. See the comparison in the left chart in Figure 2.

0.5 (3 points) Task 4b: Evaluate the performance of 1997 linear and ARIMA models

In our 1997 report, our ARIMA model projected the first time that CO2 would cross 420 ppm in May 2031. Sadly we already crossed this level in April 2021, 10 years earlier than our projection.

To quantify the model projection errors/biases, we calculated the RMSE of the models. Among the three models, the ARIMA model has the lowest RSME (8.2), in comparison to the linear model (14.1) and the polynomial model (17.6). See the comparison in the right chart in Figure 2.

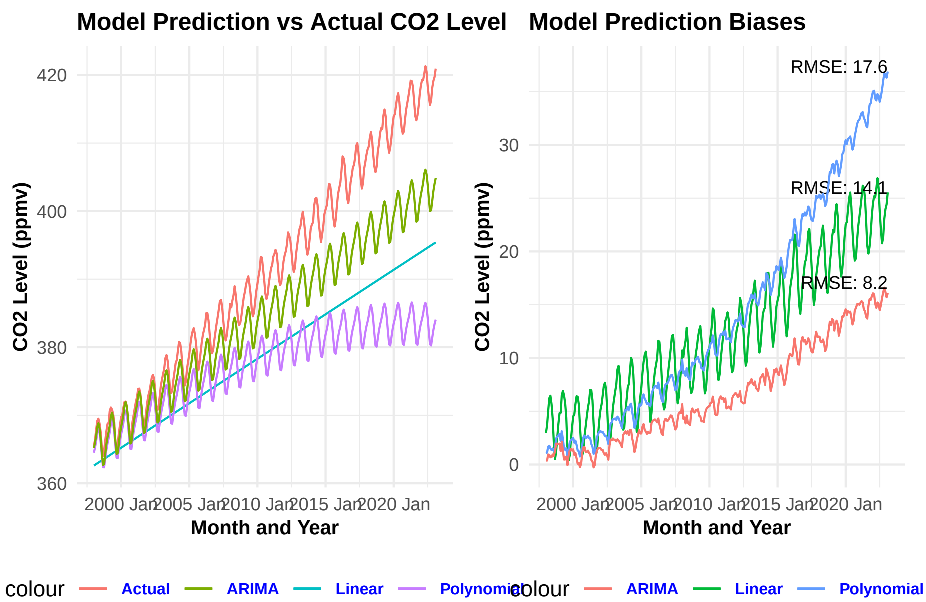

Figure 2: Model Prediction vs Actual (Left) and Model Prediction Biases (Right)

0.6 (4 points) Task 5b: Train best models on present data

While the 1997 ARIMA model outperformed the linear models, it still under-predicted actual CO_2 levels after 2005. We decided to refit the ARIMA model using the actual data.

We seasonally adjusted the weekly series, and split both seasonally-adjusted (SA) and non-seasonally-adjusted (NSA) series into training and test sets, using the last two years of observations as the test sets.

In our 1997 model fit process, we discussed the rationale of using BIC as the goodness-of-fit assessment criteria to choose the best model fit (Task 3a). We believe this process is still appropriate for refitting the ARIMA model using the SA and NSA data series. We will choose the best ARIMA models with the lowest BIC score.

Using this model optimization process, we selected the best ARIMA model for the weekly NSA data series ("ARIMA Non-Seasonal Model") as ARIMA(0,1,1)(2,1,0)[52]. This model has seasonal parameters. The best ARIMA model for the weekly SA series ("ARIMA Seasonal Model") is ARIMA(1,1,1) with drift, which does not have seasonal parameters given the data was seasonally adjusted. All coefficients for the ARIMA models are significant, see the model results below. We also fitted a third degree polynomial model to the SA data series ("Polynomial Seasonal").

```
## Series: value
## Model: ARIMA(0,1,1)(2,1,0)[52]
##
## Coefficients:
##           ma1     sar1   sar2
##       -0.7415  -0.6652  -0.33
## s.e.   0.0178   0.0199   0.02
##
## sigma^2 estimated as 0.2258:  log likelihood=-1570.7
## AIC=3149.4   AICc=3149.42   BIC=3172.41
```

```
## Series: seasonal_adj_value
```

```
## Model: ARIMA(1,1,1) w/ drift
##
## Coefficients:
##          ar1      ma1  constant
##       0.1991  -0.8308    0.0279
## s.e.  0.0255   0.0136    0.0013
##
## sigma^2 estimated as 0.1312:  log likelihood=-958.4
## AIC=1924.79   AICc=1924.81   BIC=1947.89
```

We performed diagnostic analysis of the model residuals to check for residuals stationarity and any assumption violations. The two ARIMA models have stationary residuals and a normal distribution of the residuals, with a couple of significant lags on the ACF plot. The Polynomial Seasonal model residuals are non-stationary and have significant and decaying autocorrelation lags, which require data transformation to detrend and refit.

We compared the model residuals for both the in-sample and out-sample periods for the three models, shown in Figure 3. Visually, the polynomial model has higher errors than the two ARIMA models for in-sample and out-sample periods.

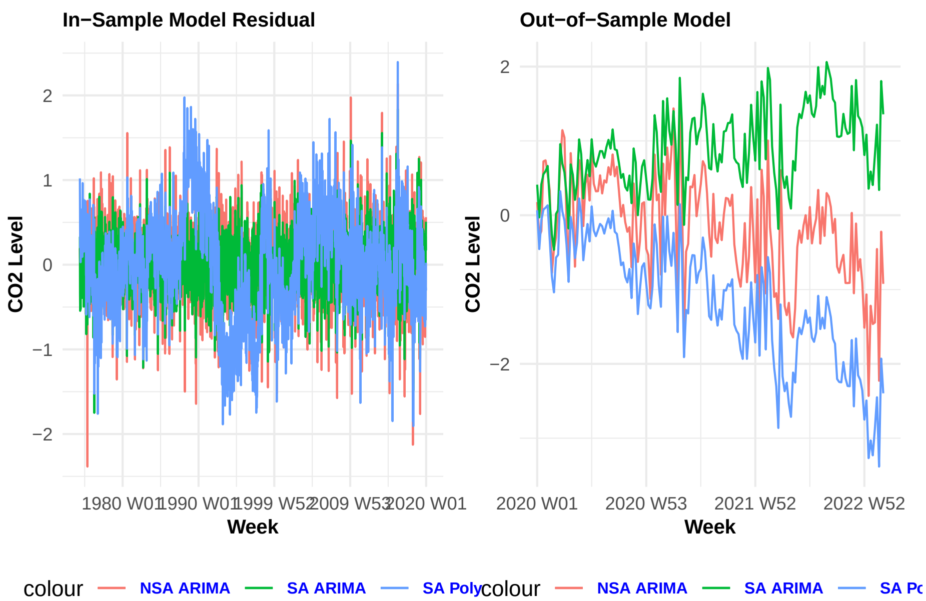

Figure 3: In-Sample Model Residuals (Left) Out-Sample Model Residuals (Right)

We calculated each model's RSME for in-sample and out-of-sample periods, see the table below. The ARIMA Seasonal model has the lowest RSME for the in-sample periods, and the ARIMA Non-Seasonal model has the lowest RSME for the out-of-sample periods. Out-of-sample period performance is more important, given the unseen data. We concluded that the ARIMA Non-Seasonal model is the best model for this data set.

model	RMSE_In_Sample	RMSE_Out_Sample
ARIMA Non-Seasonal	0.470	0.725
ARIMA Seasonal	0.362	1.033
Polynomial Seasonal	0.631	1.418

0.7 (3 points) Task Part 6b: How bad could it get?

We used the ARIMA Non-Seasonal model to predict the CO_2 levels through 2122. We projected CO_2 to reach 420 ppmv in Week 18 of 2021, which matches the actual data (420 ppmv in April 2021). Our model projects the CO_2 level to reach 500 ppmv in Week 32, 2056, and reach 670 ppmv by 2122.

As we observed from the Figure 4, the confidence interval gets wider as the forecast horizon increases. We are more relatively confident about the short term forecast, but not confident about the long term forecast.

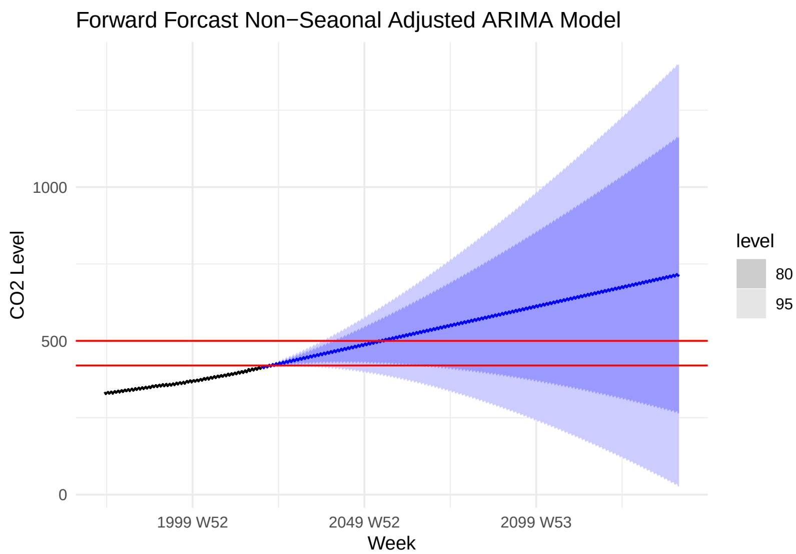

Figure 4: Non-Seasonally Adjusted ARIMA forward forecast model

Time	Forecast CO2 Level
2021 W18	420.1014
2056 W32	499.9503
2122 W01	670.7396