1 The Mauna Loa CO_2 Concentration

1.1 Problem statement

In 1958, Charles David Keeling (1928-2005) from the Scripps Institution of Oceanography began recording CO_2 concentrations in the atmosphere at an observatory located at about 3,400 m altitude on the **Mauna Loa Volcano** on Hawaii Island.

The location was chosen because it is not influenced by changing CO_2 levels due to the local vegetation and because prevailing wind patterns on this tropical island tend to bring well-mixed air to the site. While the recordings are made near a volcano (which tends to produce CO), wind patterns tend to blow the volcanic CO away from the recording site.

Air samples are taken several times a day, and concentrations have been observed using the same measuring method for over 60 years. In addition, samples are stored in flasks and periodically reanalyzed for calibration purposes. The result is a data set with very few interruptions and very few inhomogeneities.

Let C_i be the average CO_2 concentration in month i ($i = 1, 2, ...$, counting from March 1958). We will look for a description of the form:

$$C_i = F(t_i) + P_i + R_i \tag{1}$$

where:

- $F : t \to F(t)$ accounts for the long-term trend;
- t_i is time at the middle of the i^{th} month, measured in **fractions of years** after Jan 15, 1958. Specifically, we take $t_i = \frac{i+0.5}{12}$ where $i = 0$ corresponds to Jan, 1958, adding 0.5 is because the first measurement is halfway through the first month;
- P_i is periodic in i with a fixed period, accounting for the seasonal pattern;
- R_i is the remaining residual that accounts for all other influences.

The goal of is to **fit the data and understand its variations**.

Note The decomposition is meaningful only if the range of F_i much larger than the amplitude of the P_i and this amplitude in turn is substantially larger than that of R_i .

1.2 Trend estimation

In this section we fit three polynomial trends of order $n = 1, 2, 3$ to the CO_2 concentration data.

$$F_1(t) \sim \alpha_0 + \alpha_1 \quad (2)$$

$$F_2(t) \sim \beta_0 + \beta_1 t + \beta_2 t^2 \quad (3)$$

$$F_3(t) \sim \gamma_0 + \gamma_1 t + \gamma_2 t^2 + \gamma_3 t^3 \quad (4)$$

Before fitting these models, we perform 80:20 train to test partition. This way we can measure performance using RMSE and MAPE on both train and test sets, and detect potential overfitting.

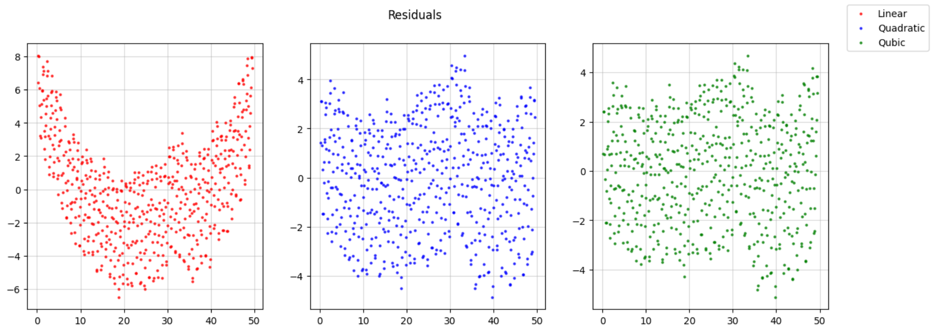

Figure 1: Residuals for each trend model

When we fit the linear model, on the residual plot we can clearly see U-shaped figure, which implies that there is a quadratic component. Residual almost doesn't change for quadratic and qubic models.

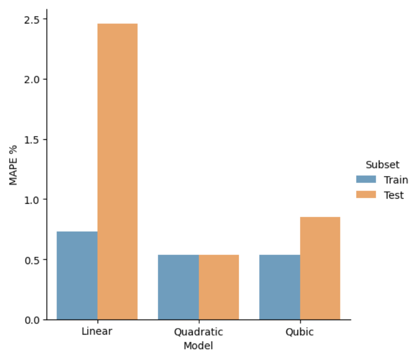

Figure 2: Model performance measured with MAPE

Also, on the MAPE bar chart, we notice that qubic model overfits the data. Thus, it's suggested to use quadratic F_2 as the trend model.

1.3 Seasonality

Let's compare ACF for C_i, and for $C_i - F_2(t_i)$ i.e. after we have removed the trend.

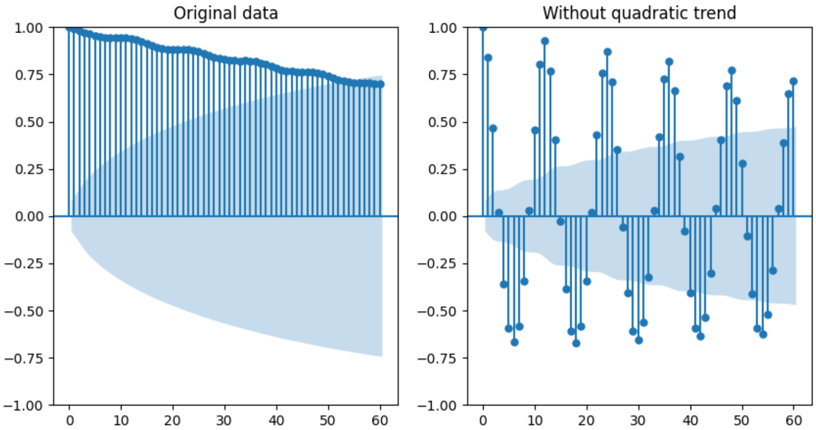

Figure 3: ACF before and after removing trend

We see that there is a clear periodic pattern on the ACF plot for $C_i - F_2(t_i)$.

The easiest way to extract the periodic component is to collect all the residuals (from removing quadratic trend F_2) for each month over the years and average them to get one data point for respective month. Then, the collection of these points can be interpolated to form a periodic signal, that is:

$$P_{i+12} = P_i \tag{5}$$

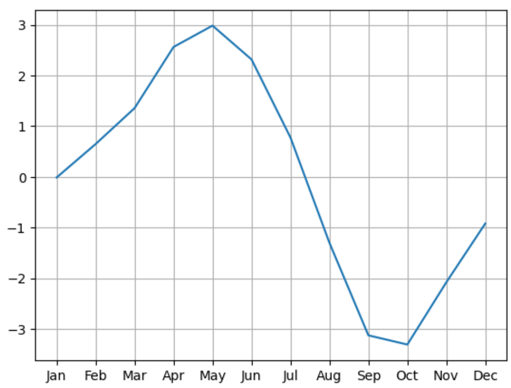

Figure 4: Periodic component of CO_2 concentration

1.4 Final decomposition and results

Comparing prediction with the test data, the prediction diverges as we go further in time.

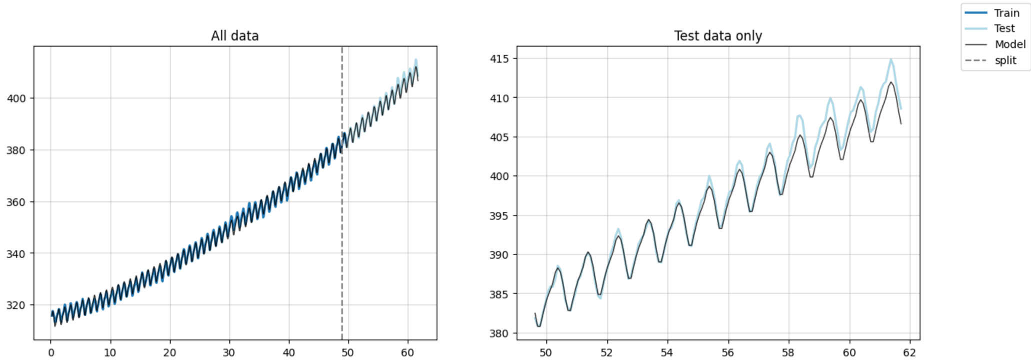

Figure 5: Real CO_2 concentration vs prediction

Final model performance measured by RMSE and MAPE is more than two times better, than for the previous model with deterministic quadratic trend F_2. This tells us the significant role of periodic component in CO_2 concentration.

Model	RMSE	MAPE
$F_2(t_i)$	2.508	0.534 %
$F_2(t_i) + P_i$	1.153	0.21 %

Table 1: Performance comparison

To show that the decomposition is meaningful, let's compare the range of values $\max - \min$ for each part of the decomposition of the training set.

Ratio	Value
F to P_i	10.99
P_i to R_i	1.64

Table 2: Ratio of amplitudes of decomposition components

We can also look at the ACF plot for our final residual. If we compare it with the previous such plot, the ACF drops to zero relatively quickly, but still doesn't fully support the stationarity hypothesis. Despite the fact we have fitted two main components (trend + seasonality), and the model seems to satisfy our definition of meaningfulness, and the mean $\bar{R}_i = -1.045 \cdot 10^{-14} \approx$

0, there could still be a **certain external regressor** which we weren't able to fit.

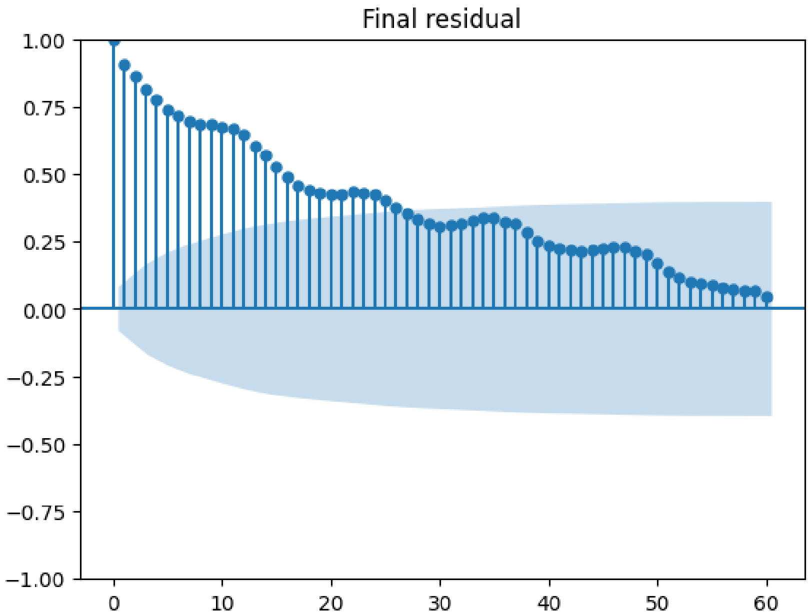

Figure 6: ACF for R_i residual