35k_file_analysis

2024-03-22

Loading the data

Currently we will be doing our analysis on the first half of the data

```
first_half <- fread("../Data/Zhu Lab First Half.csv")
# second_half <- fread("../Data/Zhu Labs Data part2.csv")
job_summary <- read.csv("../Data/Job Summary.csv")
```

Since we are only interested in the top preforming simulations, let's filter out the data; We are only considering data that: - Has data for all 5 years - The final value for Total CO2 capture is over 3.5

```
first_half.top_data <- first_half |>
  filter(year >= 5, Total_CO2_capture >= 3.5)

first_half.top_data <- first_half[first_half$file_id %in% first_half.top_data$file_id]
```

Since the *first_half.top_data* is still somewhat of a huge dataset, we will need to sample it down for preforming some basic plots and analysis.

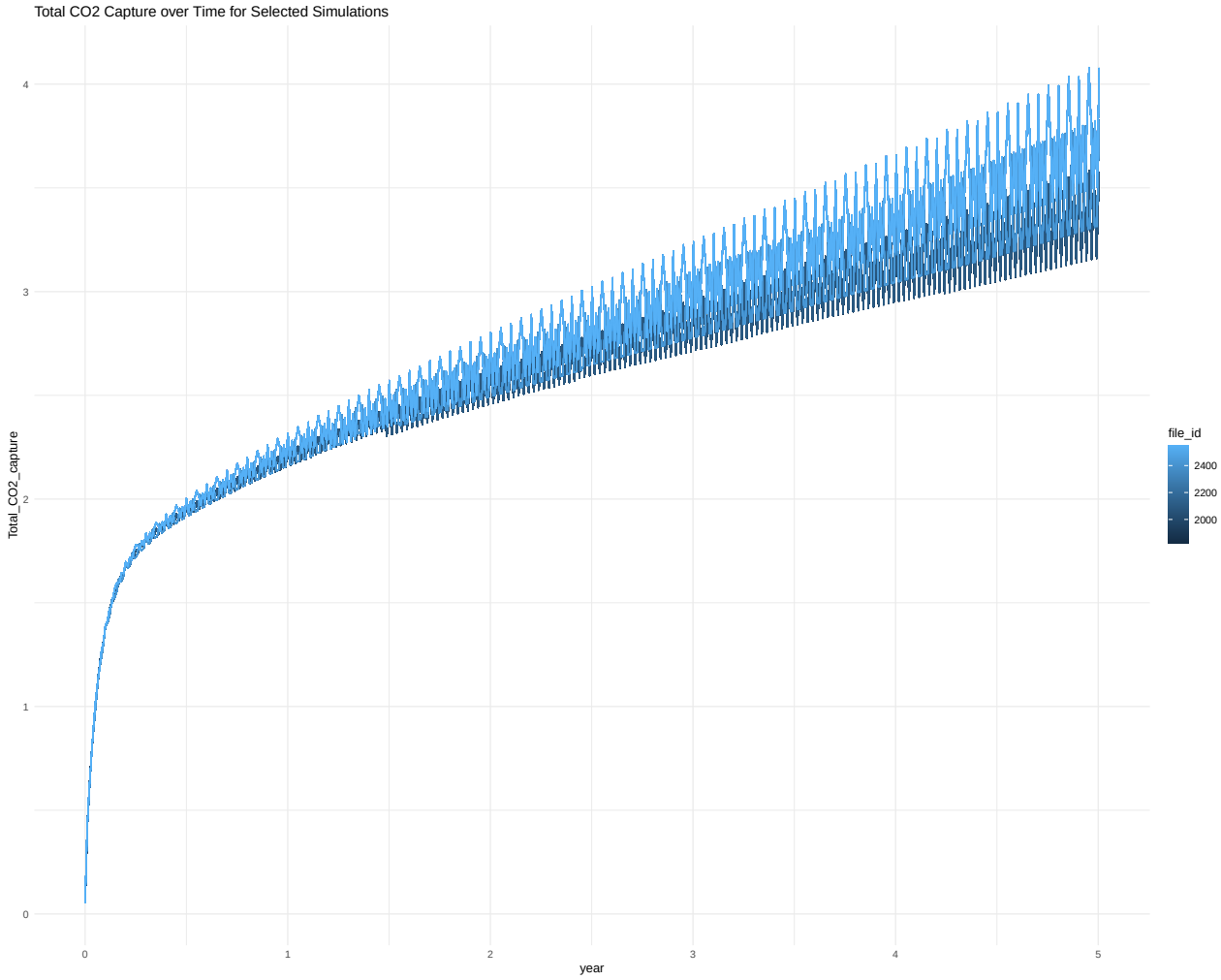
```
sample_first_half <- first_half.top_data[1:2000000, ]

sample_first_half <- merge(sample_first_half, job_summary, by = "file_id", na.rm = TRUE)

write.csv(sample_first_half, "35k_12_percent_data.csv")
```

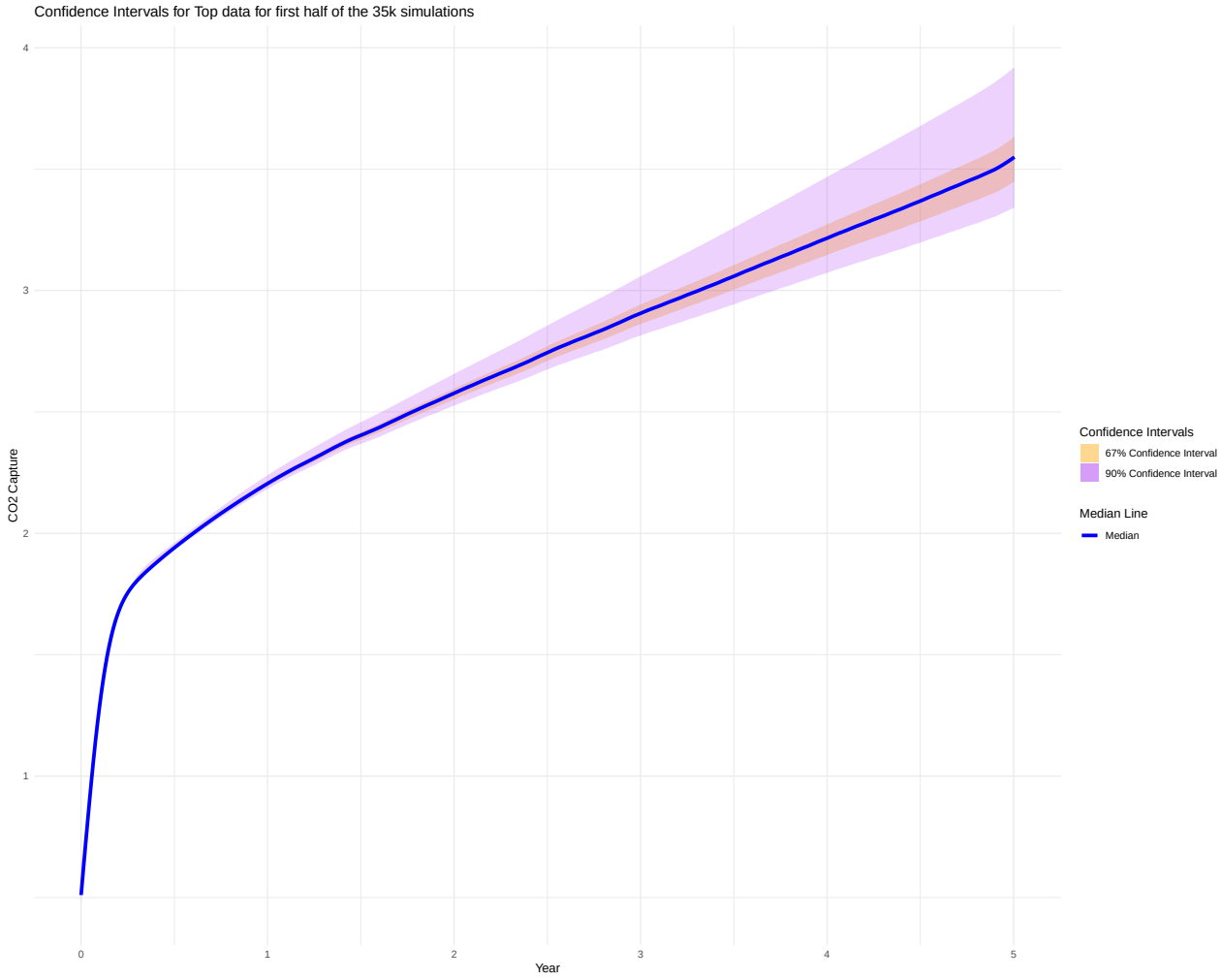
Plots

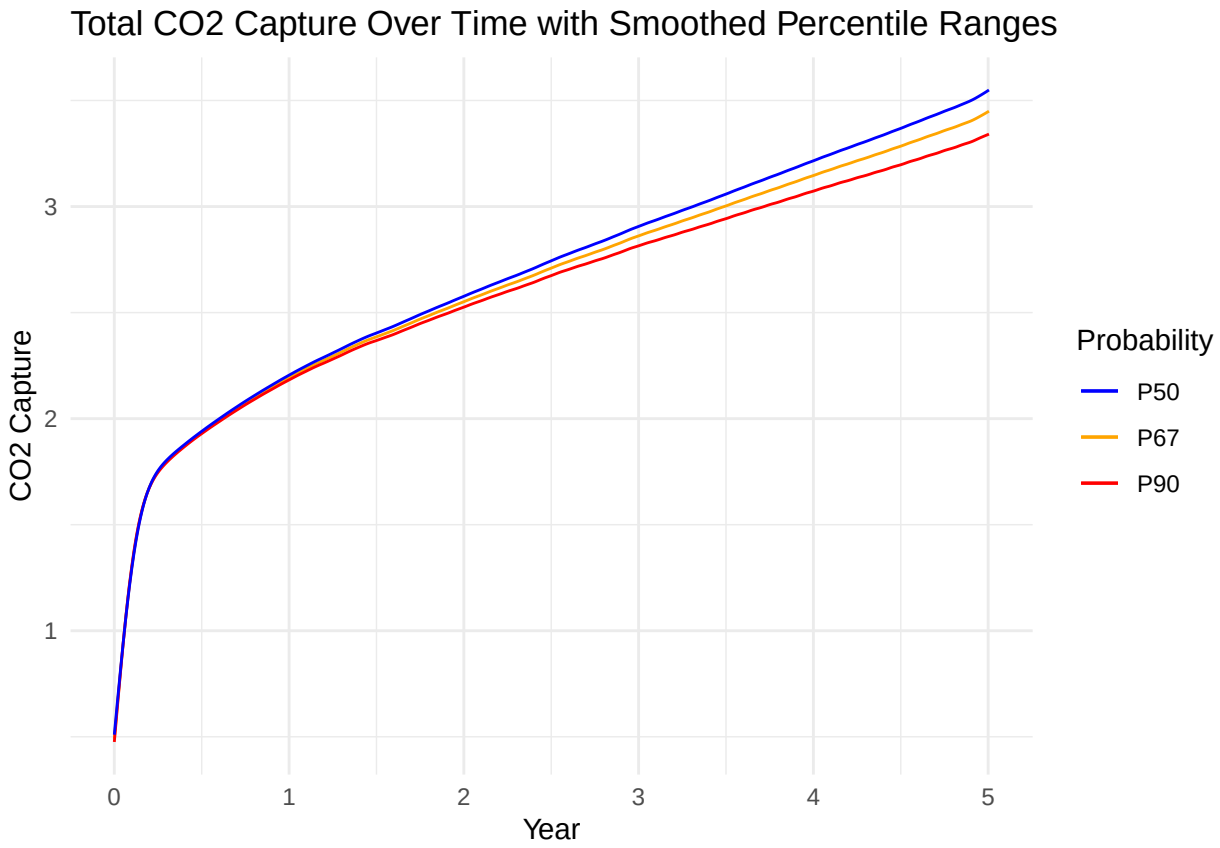
Here is a plot for the time series for different file_id, we are using the *sampled data* accounting for a total of 409 unique simulations and 2000000 rows



Quantiles

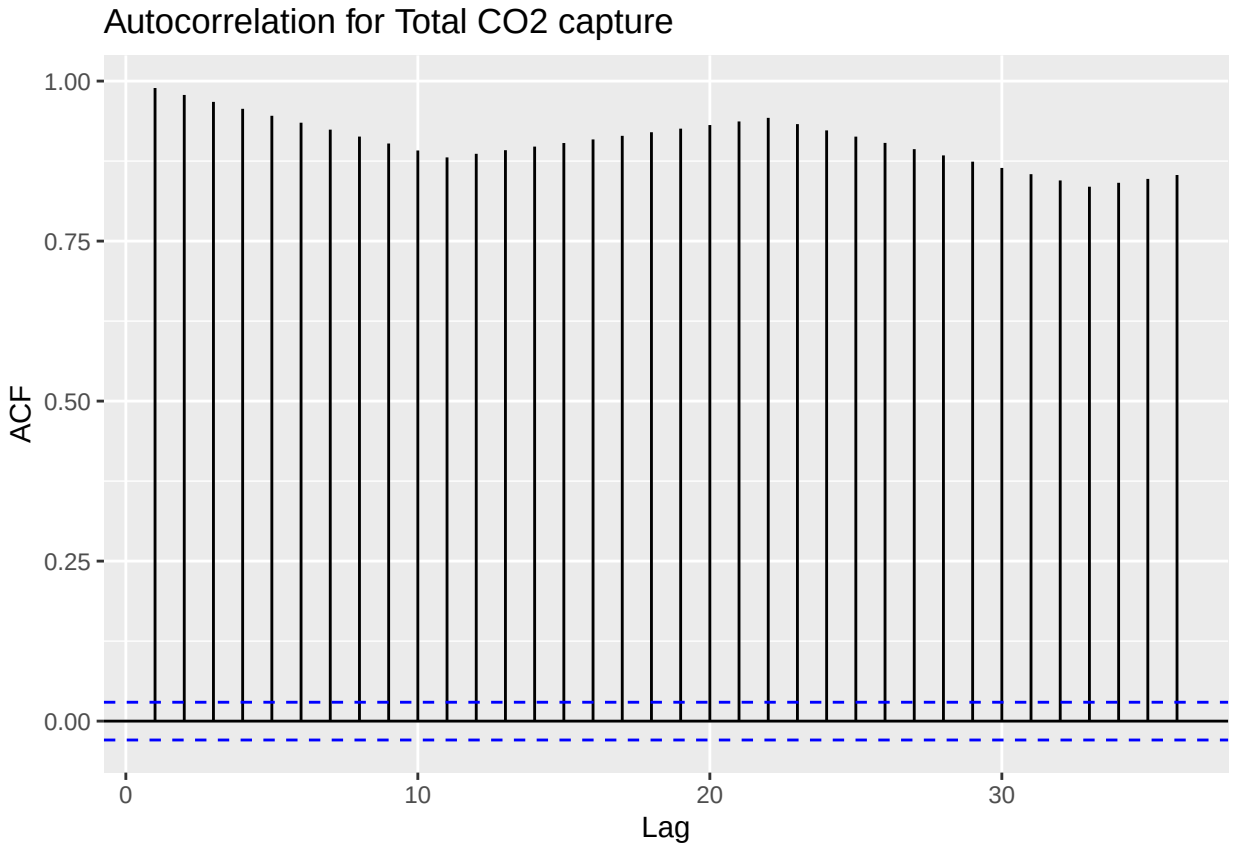
Let's also take a look over the acutal range of this data we have, we will use the entier Top Preforming simulation from the first half of the simulations for this task.

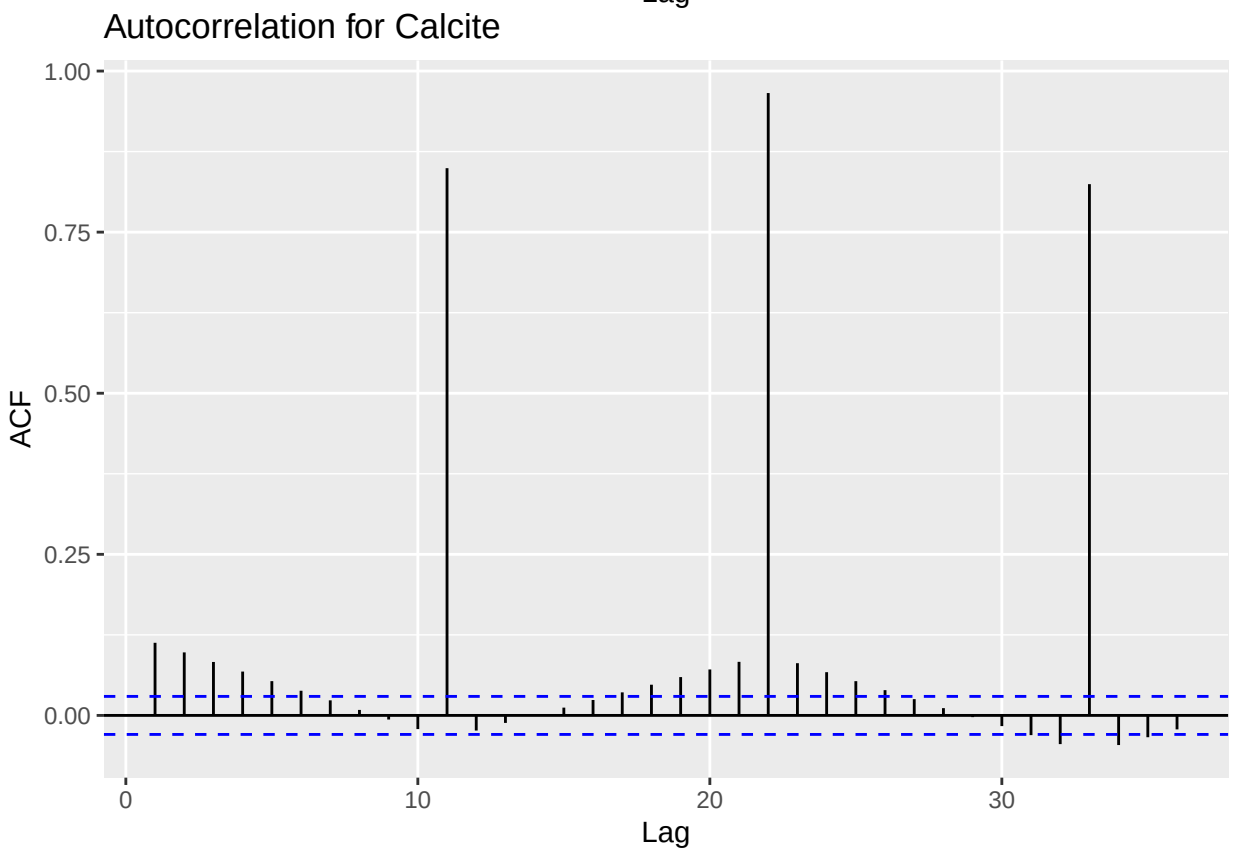

Confidence Intervals for Top data for first half of the 35k simulations
CO2 Capture
Year
0
1
2
3
4
5
1
2
3
4
Confidence Intervals
67% Confidence Interval
90% Confidence Interval
Median Line
Median

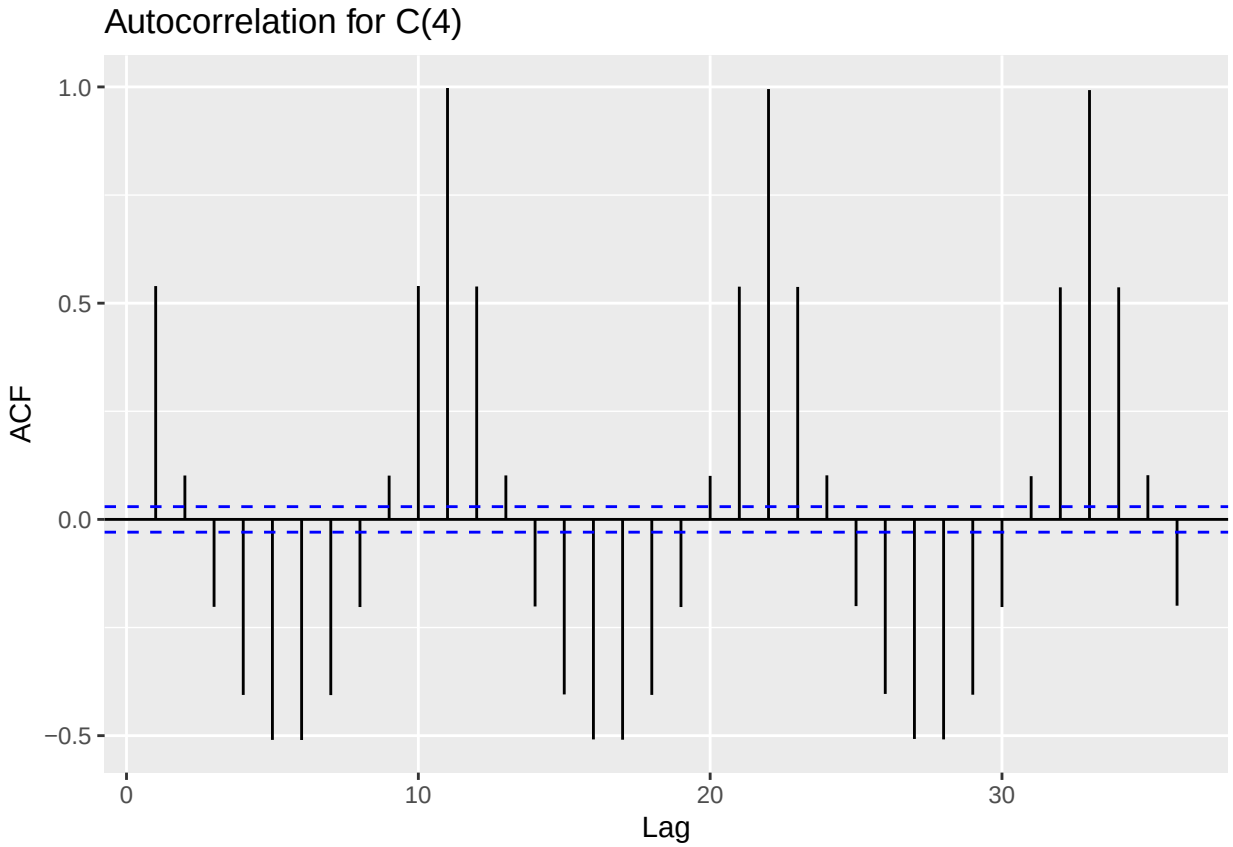


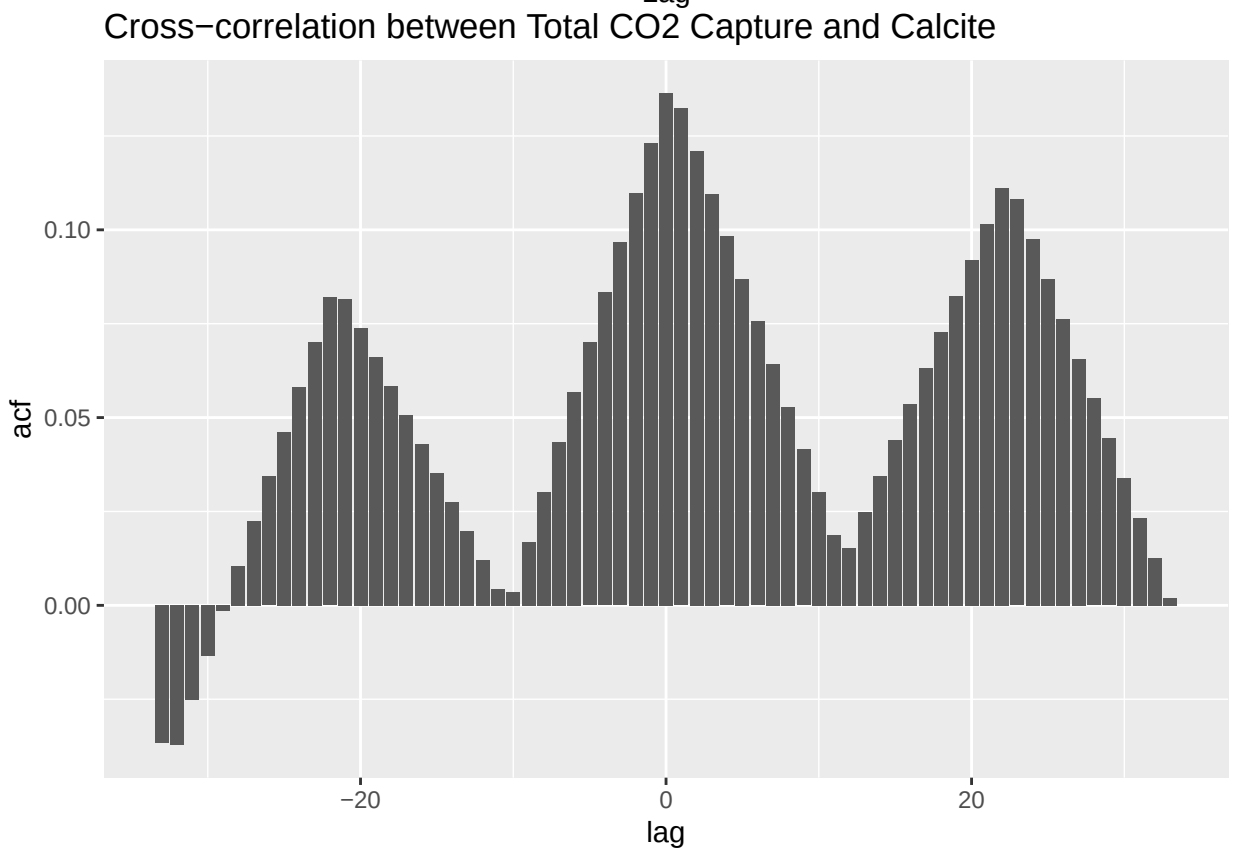
ACF, CCF

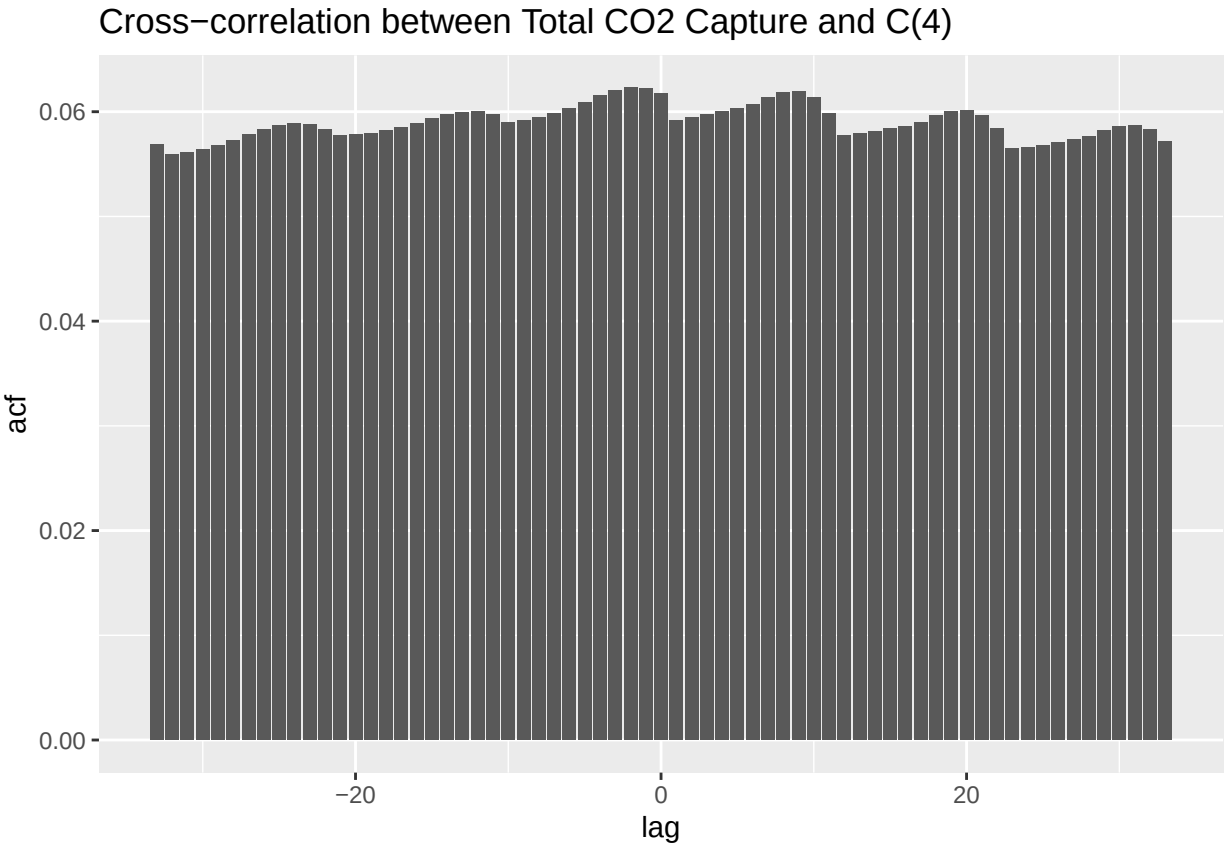
Since we can see some cyclic repersentation in the data, let's take a look at tha acf and ccf for a better look


Autocorrelation for Total CO2 capture
1.00
0.75
0.50
0.25
0.00
ACF
0
10
20
30
Lag


Autocorrelation for Calcite
1.00
0.75
0.50
0.25
0.00
ACF
0
10
20
30
Lag


Autocorrelation for C(4)
1.0
0.5
0.0
−0.5
ACF
0
10
20
30
Lag


Cross−correlation between Total CO2 Capture and Calcite
0.10
0.05
0.00
acf
−20
0
20
lag



Autocorrelation for Total CO2 Capture The autocorrelation plot for Total CO2 Capture shows strong autocorrelation across all lagged values. This indicates a high level of persistence in the time series, where past CO2 capture levels are a good predictor of future levels. Such a time series is often referred to as 'non-stationary' because its statistical properties, such as mean and variance, are not constant over time.

Autocorrelation for C(4) The autocorrelation plot for C(4) concentrations reveals a pattern that suggests possible periodicity or a repeating cycle in the time series. The alternating positive and negative lags indicate that the series oscillates over a fixed period, which could be of interest for further analysis to understand the underlying cycles in the mineral concentration data.

Autocorrelation for Calcite The autocorrelation for Calcite also shows a significant autocorrelation at lag 0 (as expected since it's a correlation with itself), with some peaks at subsequent lags. However, the peaks are not as pronounced or as regular as those for C(4), which may suggest less periodicity in the Calcite concentration data compared to C(4).

Cross-correlation between Total CO2 Capture and Calcite The cross-correlation plot between Total CO2 Capture and Calcite indicates that there are multiple time points where the correlation peaks, suggesting a relationship between the two variables. The symmetrical nature of the plot around lag 0 suggests that as one series peaks, the other series tends to peak at the same time or shortly after.

Cross-correlation between Total CO2 Capture and C(4) The cross-correlation plot between Total CO2 Capture and C(4) shows a consistent level of correlation across the lags, indicating a strong relationship between these two variables. The uniformity across lags suggests that any shifts in C(4) concentrations are consistently related to shifts in Total CO2 Capture.

Conclusions from the Analysis: - The Total CO2 Capture time series data appears to have a strong internal consistency, with current values heavily influenced by past values. - Both C(4) and Calcite show some degree of periodic behavior, with C(4) showing a more defined cycle. This could be indicative of underlying processes or seasonal effects influencing these concentrations. - There is a significant relationship

between the concentrations of Calcite and C(4) and Total CO2 Capture, which supports the hypothesis that these mineral concentrations are important predictors of CO2 capture.

Clustering Analysis