Stat 443: Time Series and Forecasting

Assignment 3: Time Series Models

Caden Hewlett

March 17, 2024

Question 1: El Niño Forecasting

The file `NINO34.csv` contains the monthly El Niño 3.4 index from 1870 to 2023. The El Niño 3.4 index represents the average equatorial sea surface temperature (in degrees Celsius) from around the international dateline to the coast of South America.

Part a

Perform exploratory data analysis.

Part a.1

Import the data into R and create a time-series object for the El Niño 3.4 index.

```
# import data
setwd('C:/Users/caden/OneDrive/Desktop/STAT_443/STAT443')
df <- read.csv("data/NINO34.csv")

# raw data is in a bad format, need to pivot it without years
dflong <- df %>%
  pivot_longer(cols = -Year, names_to = "Month", values_to = "Value")
# re-translate into dates with month abbreviations
dflong <- dflong %>%
  mutate(Date = make_date(Year, match(Month, month.abb), 1))
# then, format into a time series!
nino_ts <- ts(dflong$Value,
              start = c(1870, 1), frequency = 12)

# time series length = data frame length = 154 years * 12 months per year
stopifnot(all.equal(length(nino_ts), nrow(df)*12, 154*12))
```

Break the time series object into a training and test set. You can use the function `window()` on a ts object to split the data. Let the training set be from January 1870 to December 2021, and let the test set start in January 2022 and end in November 2023.

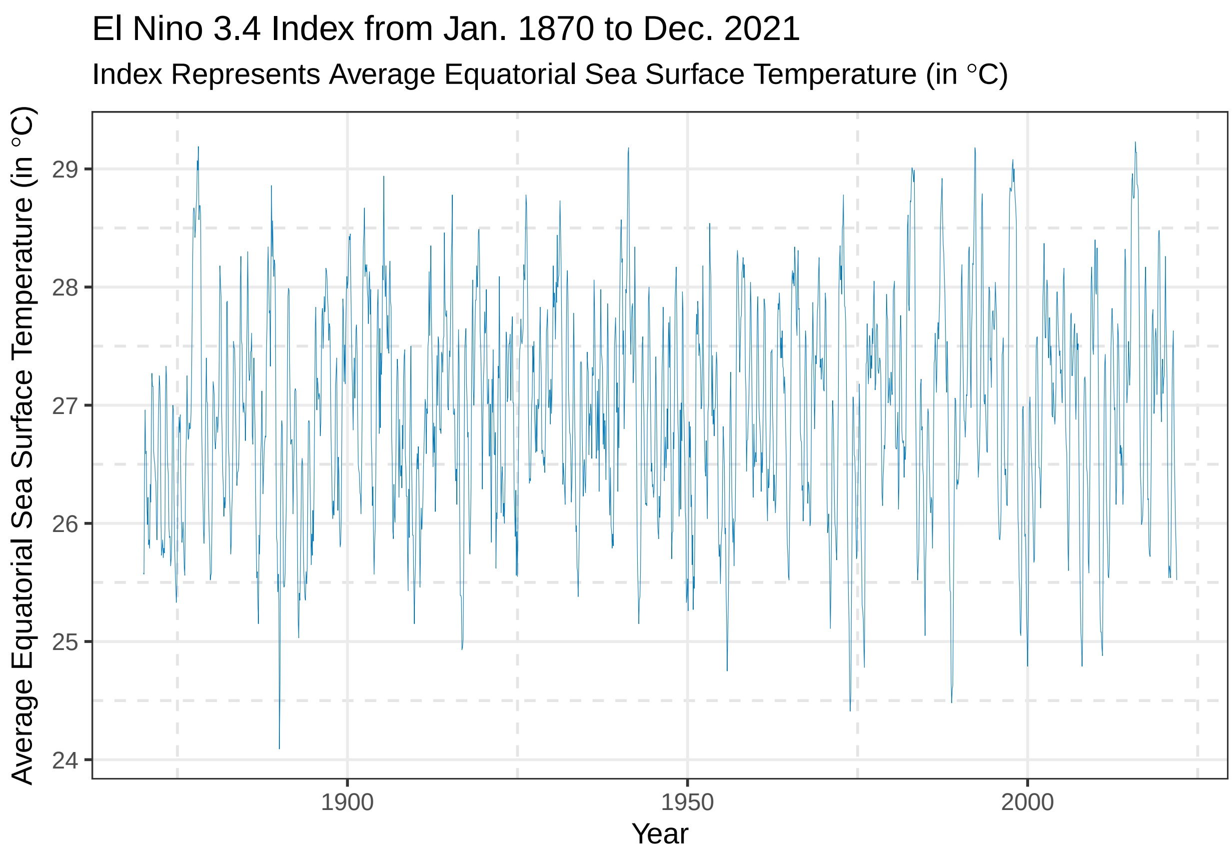
```r
# split data into train and test
nino_train = window(nino_ts, start = c(1870, 1), end = c(2021, 12))
nino_test = window(nino_ts, start = c(2022, 1), end = c(2023, 11))
# verify split
stopifnot(all.equal(
          length(nino_test)+length(nino_train), length(na.remove(nino_ts))
      )
)
```

Part a.2

Plot the training data as well as its acf and pacf.

Training Data

```r
p1data = fortify.zoo(nino_train)
p1 <- ggplot(p1data, aes(x = Index, y = nino_train)) +
  geom_line(color = "#0077b6", linewidth = 0.1) +
  labs(
    title = "El Nino 3.4 Index from Jan. 1870 to Dec. 2021",
    subtitle = expression(
      paste("Index Represents Average Equatorial Sea Surface Temperature (in ",
          degree, "C)")),
    y = expression(
      paste("Average Equatorial Sea Surface Temperature (in ", degree, "C)")),
    x = "Year"
  ) + theme_bw() +
  theme(panel.grid.minor = element_line(
    color = "grey90",
    linetype = "dashed",
    linewidth = 0.5
  ))
print(p1)
```

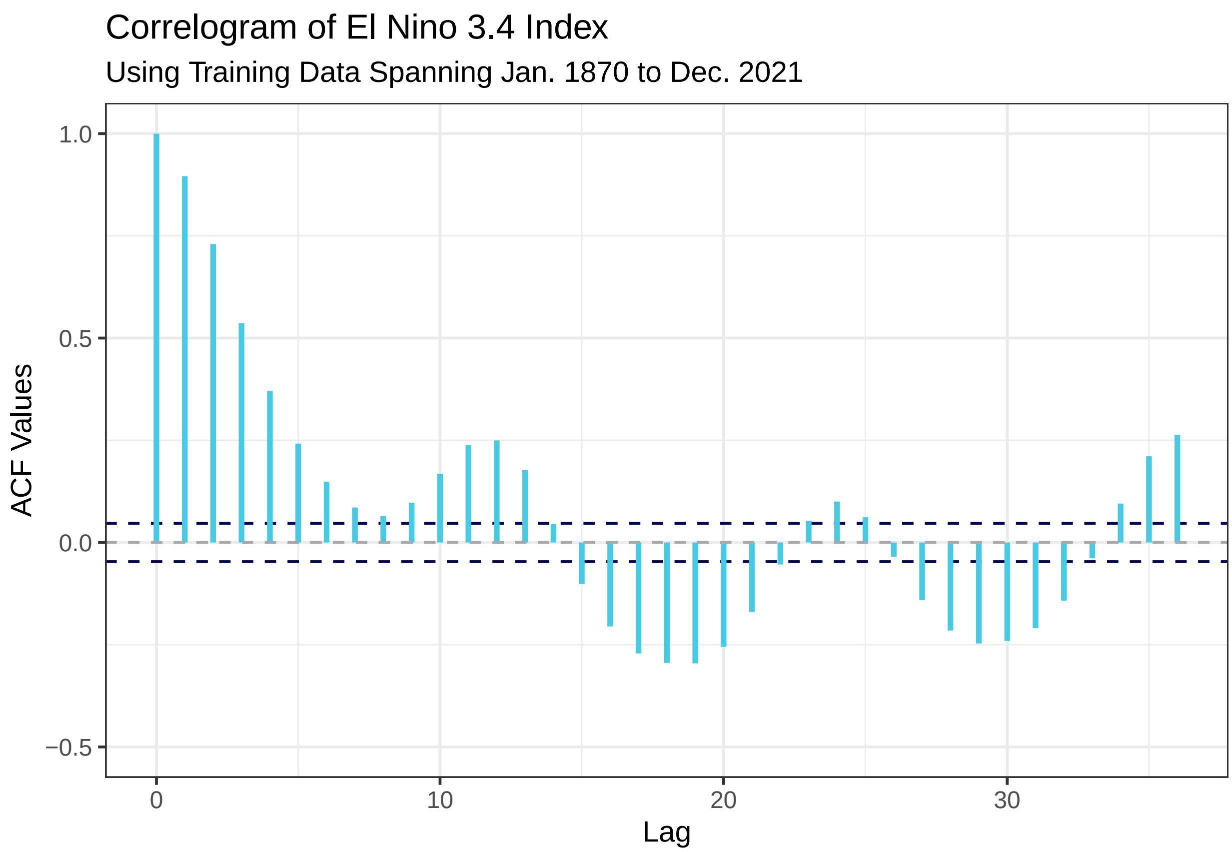


<u>Autocorrelation Function</u>

```
p2data = data.frame(
  h = 0:36,
  rh = acf(nino_train, plot = FALSE, lag.max = 36)$acf
)
n = length(nino_train)
p2 <- ggplot(p2data, aes(x = h, y = rh)) +
  geom_hline(yintercept = 2/sqrt(n),
             linetype = "dashed",
             col = "#03045e") +
  geom_hline(yintercept = -2/sqrt(n),
             linetype = "dashed",
             col = "#03045e") +
  ylim(-0.5, 1) +
  geom_segment(aes(xend = h, yend = 0),
               color = "#48cae4",
               linewidth = 1) +
  geom_hline(yintercept = 0,
             linetype = "dashed",
             color = "darkgray") +
  labs(x = "Lag",  y = "ACF Values",
       title = "Correlogram of El Nino 3.4 Index",
       subtitle = "Using Training Data Spanning Jan. 1870 to Dec. 2021") +

  theme_bw()
```

```r
print(p2)
```

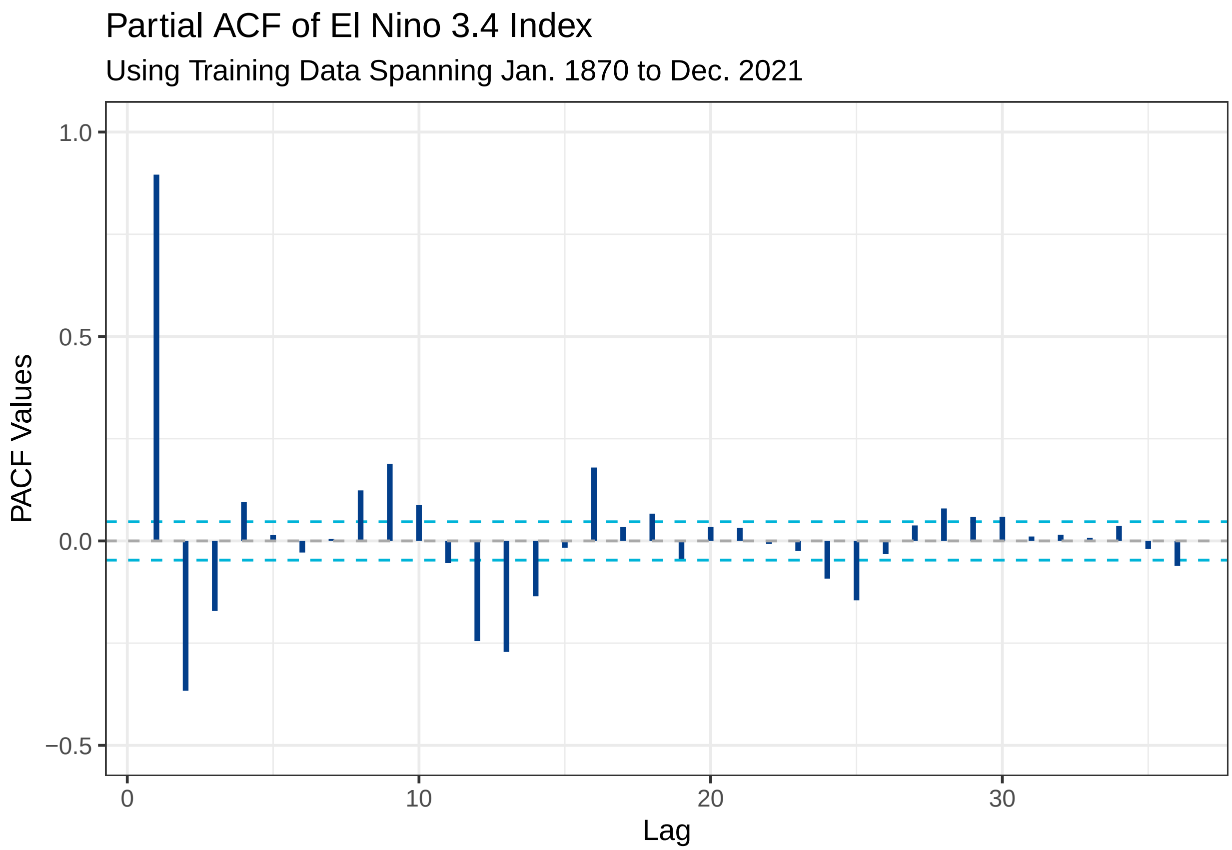


Partial Autocorrelation Function

```r
p3data = data.frame(
  h = 1:36,
  rhh = pacf(nino_train, plot = FALSE, lag.max = 36)$acf
)

p3 <- ggplot(p3data, aes(x = h, y = rhh)) +
  geom_hline(yintercept = 2/sqrt(n),
             linetype = "dashed",
             col = "#00b4d8") +
  geom_hline(yintercept = -2/sqrt(n),
             linetype = "dashed",
             col = "#00b4d8") +
  ylim(-0.5, 1) +
  geom_segment(aes(xend = h, yend = 0),
               color = "#023e8a",
               linewidth = 1) +
  geom_hline(yintercept = 0,
             linetype = "dashed",
             color = "darkgray") +
  labs(x = "Lag",  y = "PACF Values",
       title = "Partial ACF of El Nino 3.4 Index",
       subtitle = "Using Training Data Spanning Jan. 1870 to Dec. 2021") +
```

```r
  theme_bw()
print(p3)
```

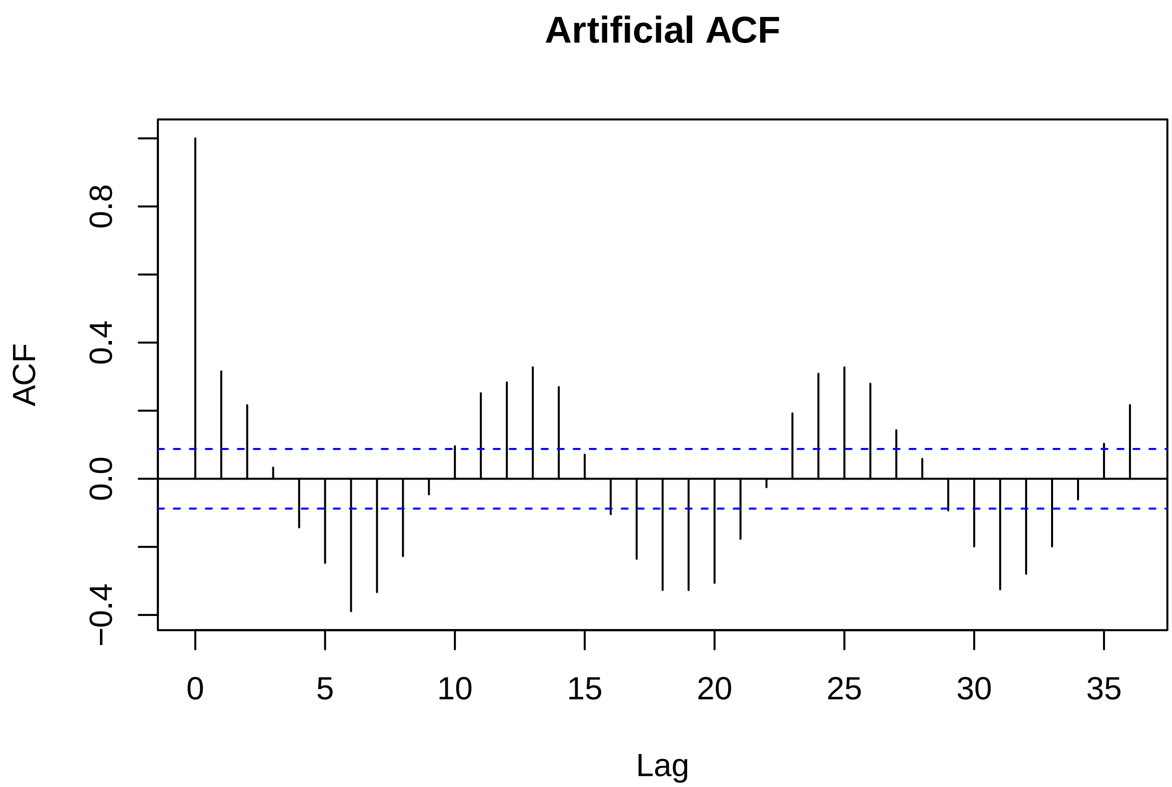


Comments

From the ACF, we see that the series exhibits a clear trend. We know that this trend is present due to the sinusoidal component of the ACF.

In fact, to replicate this pattern exactly. Let $\{Z_t\}_{t=1}^{500} \overset{\text{iid}}{\sim} N(0, 1)$ and let $X_t = Z_t + \sin(t/2)$, for $t \in \{1, 2, \ldots, 500\}$. A plot of this artificial additive seasonal model is below:

```r
Z = rnorm(500)
t = seq(1:500)
acf(Z + sin(t/2), main = "Artificial ACF", lag.max = 36)
```



As you can see, this artificial ACF closely matches our observed ACF from earlier. As such, it is very likely that there exists some seasonal term s_t. Further, from a more observational perspective, a plot of the original series data doesn't seem to support a non-constant seasonal amplitude (as in changing peak heights between seasons) which may indicate an additive seasonality rather than multiplicative.

Further, neither the plot of the data nor the ACF seem to indicate a significant trend in the data. If there were a trend component m_t, we would see some consistent change over time in the overall direction of the series beyond simple seasonality. It doesn't seem like there is anything like that in this series; however, a purely visual analysis isn't comprehensive so this doesn't mean that a trend component doesn't exist.

Finally, to determine whether or not the series is stationary, we recall the definition of a weakly stationary stochastic process. Specifically, we will consider the first property of a weak stationarity - that the mean is constant. We defined this formally in Assignment 1 previously, and was given as follows:

$$\text{Weak Stationarity Property One: } \exists \mu \in \mathbb{R} \text{ s.t. } \forall t \in \mathbb{Z},\ \mathbb{E}(X_t) = \mu$$

The presence of *either* a seasonal component or a trend causes a contradiction, as the expected value of the series becomes some function of t (and hence cannot be a constant.)

Hence, for our particular series, we can consider the following implication, letting s_t be the seasonal component:

$$\exists s_t \implies \mathbb{E}(X_t) = f(t), \text{ for some } f \implies \nexists \mu \in \mathbb{R} \text{ s.t. } \forall t \mathbb{E}(X_t) = \mu \implies X_t \text{ is not stationary. } \square$$

Where the conclusion of the implication is due to the negation of Property one. In short, due to the presence of a trend, the training time series is not stationary.