**Introduction**

We investigated the monthly sunspot data set that can be found at the Machine Learning Mastery website (<https://machinelearningmastery.com/time-series-datasets-for-machine-learning/>). This data set contains 2,820 observations of sunspots that were taken over the course of 230 years, from 1749 to 1983. The total number of sunspots fluctuates from year to year with an average cycle of 11 years. Using time series analysis techniques, we fit a SARIMA model to the data and produced forecasts of the next 11-year cycle for the data up to 1993.

**Method and Results**

We converted the sunspots data into a time series and then plotted the time series to get an initial look at any trend or seasonality present within the series. The time series exhibited seasonality but no upward or downward trend. We took the first order difference to remove any trend present and then a second difference to remove any seasonality which resulted in a stationary time series, which can be seen in Figure 3.

A picture containing histogram

Description automatically generated

Figure 1. Plot of total sunspots per year

Chart, line chart

Description automatically generated with medium confidence

Figure 2.

A picture containing text, antenna

Description automatically generated

Figure 3.

From the stationary time series, we were then able to examine the ACF and PACF plots to determine appropriate orders for the auto-regressive and moving average parts of the ARIMA model, for both the seasonal and non-seasonal components.

Chart, histogram

Description automatically generated

Figure 4. ACF and PACF plots of the sunspots time series.

For the seasonal components of the ARIMA model, we found that there were non-zero lags at seasonal lags 11 and 22 in the ACF plot, while the PACF plot showed only 1 convincing non-zero lag at seasonal lag 11. This suggests seasonal orders of Q = 2 and P = 1 for the seasonal component of the ARIMA model.

For the non-seasonal components, we found there to be 7 non-zero lags in the ACF plot while the PACF plot showed 8 non-zero lags. This suggests orders of q = 7 and p = 8. We took one difference for the trend and one difference for the seasonality, therefore D = d = 1 resulting in the following model: ARIMA(8,1,7)X(1,1,2)\_1.

After deriving the ARIMA model, we next evaluated the fit of the model by examining the ACF and PACF plots of the residuals among other plots, including the Ljung-Box statistic. We also viewed the QQ plot to determine whether or not any normality assumptions had been violated.

Chart, histogram

Description automatically generatedDiagram

Description automatically generated

Chart, line chart

Description automatically generated

Figure 5. Model diagnostics plots including the ACF and PACF of the residuals, plots of the standardized residuals, p-values for the Ljung-Box statistic, and the QQ plot of normality.

If the ACF and PACF plots of the residuals do not show any auto-correlation then we can say with confidence that the residuals are random white noise terms. This suggests that the model is an adequate fit for the sunspots time series. From the model diagnostics plots, we can see that there is no auto-correlation in the ACF and PACF plots, and that the normality assumptions are not violated, therefore the model is adequate.

With an appropriate model for the sunspots time series, we were then able to predict the number of sunspots for each year during the next solar cycle using forecasting. The results of this forecast can be seen in Figure 6.

Graphical user interface

Description automatically generated with medium confidence

Figure 6. Forecast of sunspots time series for the next solar cycle of 11 years.

**Discussion**

In the initial analysis, the seasonality contained in the data did not appear to be monthly dominated. As a result, the data were transformed to yearly counts prior to conducting our time series analysis, which allowed us to focus on the cyclical nature of solar cycles.

**Appendix**

knitr::opts\_chunk$set(echo = T, warning = F, message = F, fig.align = "center")

# Read libraries

library(TSA)

library(ggplot2)

library(tidyverse)

library(forecast)

library(astsa)

# Read sunspot data from csv file

sun <- read.csv("monthly-sunspots.csv", header = TRUE)

head(sun)

# Create and plot original time series

sun.ts <- ts(sun$Sunspots, start = c(1749,1), deltat = 1/12)

plot(sun.ts, xlab = "Year", ylab = "Sunspots", main = "Sunspots per Year")

# Difference time series. May be unnecessary as the data does not seem to include a trend

sun.diff <- c(NA, diff(sun.ts))

sun.diff <- ts(sun.diff, start = c(1749,1), deltat = 1/12)

plot(sun.diff, xlab = "Year", ylab = "First Order Differenced Series")

# Plot ACF and PACF of first order differenced time series

par(mfrow = c(1, 2)) # I think this is the only thing I've changed lol

acf(sun.diff, lag.max = 150, na.action = na.pass,

main = "ACF for differenced series")

pacf(sun.diff, lag.max = 150, na.action = na.pass,

main = "PACF for differenced series")

# Apply seasonal differencing. According to NASA, a solar cycle is approx 11 years. Apply a difference of 132 months lag to data.

sun.diff2 <- c(NA, diff(sun.diff, lag = 132))

sun.diff2 <- ts(sun.diff2, start = c(1749,1), deltat = 1/12)

plot(sun.diff2)

# Plot ACF and PACF for differenced series

par(mfrow = c(1, 2))

acf(sun.diff2, lag.max = 150, na.action = na.pass,

main = "ACF for Differenced Series")

pacf(sun.diff2, lag.max = 150, na.action = na.pass,

main = "PACF for Differenced Series")

# Create arima (MA) model based on ACF Cutoff and PACF tail off

n <- length(sun.ts)

fit1 <- arima(sun.ts, order = c(0,0,7), seasonal = list(order = c(1,1,1), period = 12))

fit1

# Explore residuals and fit of our model

par(mfrow = c(1,2))

res <- fit1$residuals

acf(res, lag.max = 50)

pacf(res, lag.max = 50)

tsdiag(fit1)

# Check normality assumption using QQ plot

qqnorm(res)

qqline(res)

# Since the data does not appear to follow any sort of monthly trend, convert the data to yearly counts and analyze the time series

sun.tidy <- sun %>% separate("Month", into = c("Year","Month"), sep = "-")

sun.tidy2 <- sun.tidy %>% group\_by(Year) %>%

summarise(total\_sunspots = sum(Sunspots))

head(sun.tidy2)

# Plot yearly time series data

sun.tidy.ts <- ts(sun.tidy2$total\_sunspots, start = c(1749,1), deltat = 1)

plot(sun.tidy.ts, xlab = "Year", ylab = "Sunspots", main = "Sunspots per Year")

# Apply first order difference

sun.tidy.diff <- c(NA, diff(sun.tidy.ts))

sun.tidy.diff <- ts(sun.tidy.diff, start = c(1749,1), deltat = 1)

plot(sun.tidy.diff, xlab = "Year", ylab = "First Order Differenced Series")

# Apply second order difference to remove seasonality (11 year solar cycle)

sun.tidy.diff2 <- c(NA, diff(sun.tidy.diff, lag = 11))

sun.tidy.diff2 <- ts(sun.tidy.diff2, start = c(1749,1), deltat = 1)

plot(sun.tidy.diff2, xlab = "Year", ylab = "First Order Differenced Series")

# Plot ACF and PACF

par(mfrow = c(1, 2))

acf(sun.tidy.diff2, lag.max = 50, na.action = na.pass,

main = "ACF for Differenced Series")

pacf(sun.tidy.diff2, lag.max = 50, na.action = na.pass,

main = "PACF for Differenced Series")

# ACF and PACF tail of suggesting an ARIMA model may be a good fit for our time series

n <- length(sun.tidy.ts)

fit1 <- arima(sun.tidy.ts, order = c(8,1,7), seasonal = list(order = c(1,1,2), period = 1))

fit1

# Identify model fit using residuals

par(mfrow = c(1,2))

res <- fit1$residuals

acf(res, lag.max = 50)

pacf(res, lag.max = 50)

tsdiag(fit1)

# Check the normality assumption

qqnorm(res)

qqline(res)

# Predict the number of sunspots for each year during the next solar cycle

pred <- predict(fit1, n.ahead = 11)

plot(sun.tidy.ts, xlim = c(1749, 2010), ylim = c(0, 2500), main = "Forecast of Sunspots", ylab = "Number of Sunspots")

lines(pred$pred, col = "red")

lines(pred$pred - 2 \* pred$se, col = "blue")

lines(pred$pred + 2 \* pred$se, col = "blue")