Analyzing and Forecasting The Time Series of the U.S. Mexico Border Entries Through SARIMA Model and Spectral Analysis

Phuc Lu

UC Santa Barbara

pdlu@ucsb.edu

Abstract

A time series and spectral research were conducted on a data set containing monthly U.S.-Mexico border entries from 1996 to 2024. The Box-Jenkins approach was followed. Plotting the data revealed that the number of people entering the U.S. through the U.S.-Mexico border peaked in 1999 and fell after 2000 until it rose again after 2011. The number dropped to an all-time low during the COVID-19 pandemic and seems to be correcting itself a couple of years later. A SARIMA(2, 1, 0)(2, 0, 0) was fitted to the data and forecast the number of entries for the next 12 months. The model had a low AIC, but there were some seasonal components not accounted for, which hurt the model's performance a bit. Spectral Analysis was also conducted to examine the time series through the frequency domain for potential seasonality and test for significance. It was found that there were some frequencies with strong power suggesting seasonality. Confidence intervals were created for each frequency with a high spectrum to determine its significance. It was determined that there was no true periodic signal. The work serves as an introduction to applying time series analysis and spectral analysis techniques to a time series data set and studying the analysis result. More in-depth and different techniques can be applied to build upon this work in the future.

Introduction

When it comes to immigration, the U.S. has a long history of passing laws that increase and decrease the number of immigrants. Starting in 1965, the U.S. passed immigration laws that permitted Northern and Western European immigrants to enter, while people from Asia were entirely barred from entry. The immigration policy was loosened in 1965 with the Immigration and Nationality Act which allowed people from Latin America and Asia to immigrate. The policy was signed by US President Lyndon B Johnson. This act lifted the national origins quota and capped the number of visas at 290,000 per year. This is also the first policy to limit western hemispheric immigration to the US. This made it so that it would be easier for people from Asian countries to immigrate to the US while making it more difficult for Latin Americans and Europeans to immigrate to the US. However, this also led to the invention of illegality, which made it a crime to enter the US illegally. The Immigration Act of 1990 further increased legal immigration and allowed even more immigrants from more countries to enter the U.S. legally (Moslimani & Passel, 20204).

As of today, the U.S holds more immigrants than any other countries in the world. The PEW Research Center found that immigration today accounts for 14.3% of the U.S population. This is a massive growth from 4.7% since 1970. However, the PEW Research Center in 2022 found that 23% of all U.S immigrants were born in Mexico, making Mexico the top country of birth for U.S immigrants (Moslimani & Passel, 20204).

With this in mind, the research will be focusing on analyzing U.S. border entry times data from the U.S Mexico Border. Key points of interest include analyzing trends in border entries over some time, especially during major events like the historical U.S. immigration laws, 9/11, and the COVID-19 pandemic. This project will also explore whether there are seasonal patterns or recurring cycles in the data, such as any four-year trends tied to election years. The purpose of this research is to see how an ARIMA time series model would fit the data and to see if conducting a Spectral Analysis would reveal any insight into seasonality.

Data

Initially, this project was going to be an analysis of NVIDIA stock price, but this idea was somewhat basic. The data searching process ultimately led to the discovery of a dataset on U.S. border entries from the Data.org website. Given the current political climate, with border security and immigration being highly debated topics, this provides an interesting opportunity to gain insights from the time series. I expect to see some major dips in the time series at specific times. In particular, the time around Trump and Biden's Presidency. It would be interesting to see U.S. entry trends during disaster times such as 9/11 and COVID-19. It's also important that I gain experience from working with time series data and be able to add this project to my resume.

The time series data was collected by the U.S. Customs and Border Protection (CBP) at each port of entry. The website didn't specify, but it can be assumed that data was collected through documentation of each entry by the CBP personnel. The data reflect those entering the U.S., not those leaving the U.S., and does not reflect undocumented entries. The data set starts in January 1996 and ends in December 2024. The frequency of the data set is monthly. The column names are port name, state, port code, border, date, measure, value, longitude, latitude, and point. The port name is a string for the port of entry's name and there are 116 unique ports. Port codes are a series of identifying codes for each of the ports of entry. State is a string for state names and there are 14 states. The date represents the date of entry and is encoded as a string. Value is the number of people that have entered in a particular month. The measure is a string representing how the people entered the US. Longitude and Latitude are the coordinates and point is a coordinate representation of both longitude and latitude, i.e. Point(x,y). There are 397,909 observations in the original data set. After getting entries from the U.S.-Mexico Border only, there were 93,845 observations remaining in the data set.

Methodology

Since the most important variables for this time series analysis were date and value, the entire data set was grouped by the dates variable and the summed number of people entering the country, given the value for each date. The data set is then converted into a time series object through the ts() function from the tseries library with frequency = 12 to specify that the data is monthly data.

The Box-Jenkins approach was followed to build SARIMA models. First, the time series was plotted with plot.ts() . No transformation was necessary. Stationality is checked by plotting the data, plotting the ACF, and applying the augmented Dickey-Fuller Test. The data is then differenced at lag 1 to make it stationary for further analysis. The time series data is put into the auto.arima() with seasonal = TRUE to find the appropriate model. Estimation of parameters was also observed. For model diagnosis, the following visualizations were closely examined: standardized residual plot, ACF- Residual, the Normal Q-Q Norm for standardized residuals, and the p-value for LJung-Box statistics. Model selection was chosen based on the lowest AIC value. Forecasting is made through the sarima.for(dt, n.ahead = 12) with the proper (p,d,q)(P, D, Q)[m] value given by the auto.arima() function.

Spectral Analysis was also conducted to seek out potential seasonality trends in the time series in the frequency domain. First, the data is put into the mvspec function and returns a raw periodogram. The periodogram is examined for any spectrum peaks. A smoothed periodogram with taper = 0.1 and span = c(7,7) was also implemented to better visualize the spectrum spikes. After identifying the frequencies with the highest spectrum, the Chi-Squared statistics were used to see if the cycles were significant. Confidence intervals were also produced.

Results

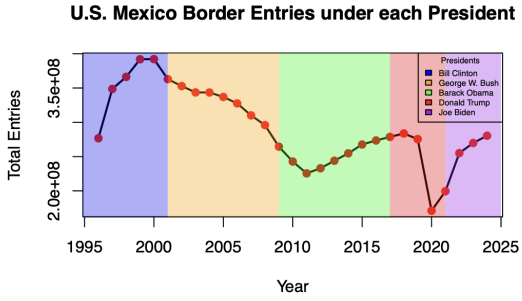


Figure 1: This plot shows the number of people entering the U.S through the U.S. Mexico border entry from 1996 to 2024 under 4 different presidencies, shown through 4 different colors.

Six before 1996 was when the U.S. passed the Immigration Act of 1990 where the U.S. loosened its policy on immigration allowing more people to enter the country. This probably explains the rising trend leading up to 1999 and 2000 when the number of people entering the U.S. through the U.S.-Mexico border peaked. This peak was also under the Clinton administration. After Bush took office, the number of U.S. border entries showed a decline throughout all 4 years under Bush. The downward trend resumed and hit a local minimum in 2011. This was probably due to 9/11. After 2011, the number of entries started to increase again slowly. Once Trump took office, there was a decreasing number of entries. Once the COVID-19 pandemic hit, there was a steep fall in the number of border entries. The year 2020 has the lowest number of entries through the U.S.-Mexico Border in the past 28 years, making this an outlier value. The number began to rise again after President Biden took office in 2021.

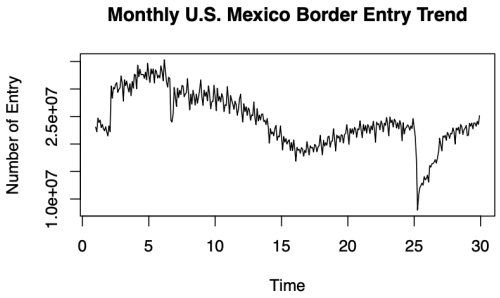


Figure 2: The time series plot below is the number of U.S entry through the U.S. Mexico border time series broken down into monthly data over the past 28 years. The time series starts at time 0, which means on January 1996. The data ends on year 28 which is December 2024.

When plotting the time series from a monthly perspective, there are a lot of fluctuations. In addition, there are consistent peaks and falls in the series. This is hinting that there might be some seasonal effects. Although from the yearly perspective, 2019 to 2020 shows a great dip in the number of entries during COVID-19. However, the monthly perspective shows one single month where the number of entries suddenly cut to reduce the amount of people traveling during the COVID-19 lockdown. In terms of the time series itself, it is not stationary and differencing is necessary.

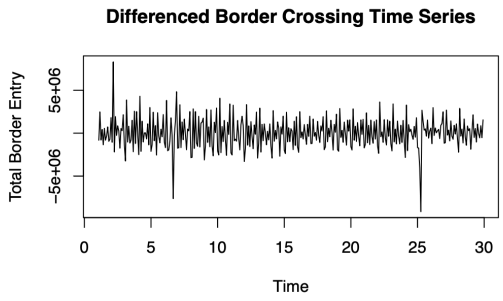


Figure 3: The time series becomes stationary after taking the first difference.

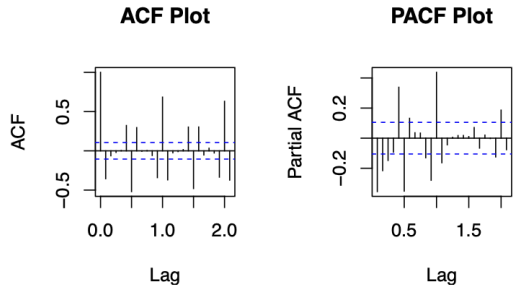


Figure 4: These two figures show the ACF and PACF Plot of the Entry through U.S. Mexico Border Time Series

The time series becomes stationary after taking the first difference. Differencing also helps to reveal 3 outlier values.

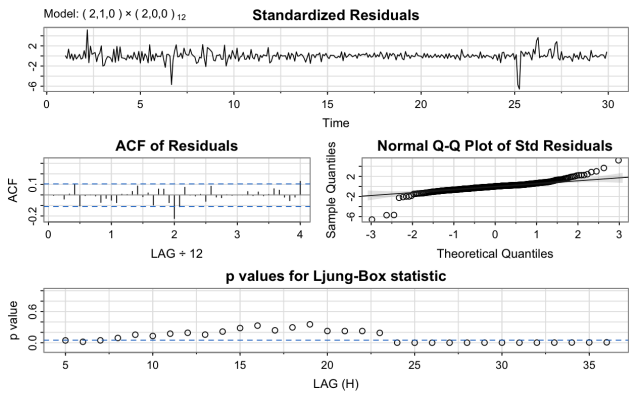
The ACF measures the correlation between the time series and its lagged values. For lag 1 to 3, the ACF shows a negative correlation. This means that there is a weak negative short-term relationship. For lag 6, there is a positive correlation indicating some positive relationship between observations that are 6 periods apart. At lag 12, there is a strong correlation suggesting some positive relationships between observations that are 12 periods apart. At lag 24, there is a strong negative relationship which indicates some long-term negative relationship between observations that are 24 periods apart. Where there are larger positive or negative values at specific lags, these suggest that there is possible seasonality. The PACF measures the correlation between the time series and its lagged values after removing the effect of intermediate lags. For lag 1 to 4, the PACF values are negative, which suggests some negative correlation between immediate past values and the current values. For lag 5, the correlation is highly positive. This means that there could potentially be a structure where past values separated by 5 periods influence the current value. At lag 12, the PACF is very much strongly positive. This suggests that there could potentially be an influence of past observations 12 periods apart on the current value. For lag 24, there is a significant positive correlation which suggests seasonality.

```
Coefficients:
          Estimate          SE t.value p.value
ar1         0.0000      0.0562 -0.0003  0.9998
ar2        -0.1358      0.0544 -2.4983  0.0129
sar1        0.4150      0.0500  8.2999  0.0000
sar2        0.3854      0.0537  7.1814  0.0000
constant 54497.4519 239004.0386  0.2280  0.8198

sigma^2 estimated as 1.436325e+12 on 342 degrees of freedom

AIC = 30.89784  AICc = 30.89835  BIC = 30.9644
```

The SARIMA model that the auto.arima() function has chosen is SARIMA(2, 1, 0)(2, 0, 0)[12]. The parameter estimates are shown in the table, which can be useful for prediction. The AIC is 30.89784 which is an acceptable value because it's a relatively low value.



There are many values that are well beyond 3 standard deviations away from 0. These are outlier values. The overall shape of the standard residuals seem to have two prominent spikes on each end of the time range. The standardized residuals on the two ends of the time frame tend to be farther from 0 on the two ends compared to the middle section. This makes sense because there was more border entry volatility during the Clinton and Bush presidency (approx. from 1996 to 2000) and around the COVID-19. In the plot of estimated ACF of residuals, all of the residuals have minimal dependency except for lag 24, where residual ACF is strongly negative. This could mean that the model didn't capture some seasonality associated with this lag. The Normal Q-Q plot shows that the assumption of normality is fairly reasonable, though there are some outliers at the ends. Q-statistic is mostly not statistically significant for the middle lags. However, the p-values at the beginning are significant, but they eventually are not significant for the middle mad times . However, the p-values are significant again after lag 23. This means that the SARIMA(2, 1, 0)(2, 0, 0)[12] is good for capturing dependencies for the middle lags from 8 to 23, but it struggles for lag 24 and beyond.

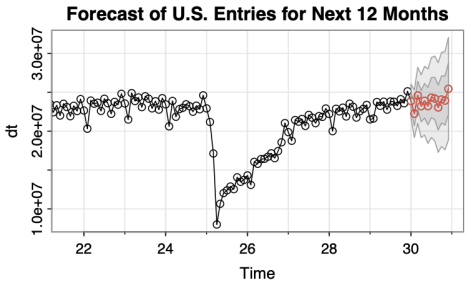
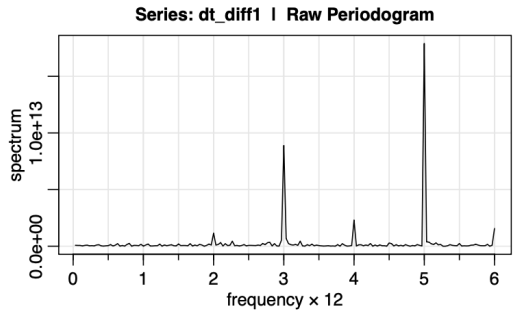


Figure 6: This chart shows the forecasted total number of people entering the U.S. through all 28 ports of entry on the U.S. Mexico border for the next 12 months after December 2024.

Visually, the predictions look great. The predicted values seem to match the values before COVID-19. It seems like the number of entries is to correct itself after the pandemic to match the numbers before. However, the standard error is a bit alarming since the predicted number of people entering the country is off by about 2 million people per month on average across all 28 ports. In addition, it is important to note that for the next 12 years starting on January 25, President Trump will take office. Because a new administration is in office, there could be some unforeseen factors such as policy changes that may affect the number of entries from Mexico.

Spectral Analysis



The dominant peak is at frequency 5 followed by 3. There are smaller peaks 2, 4, and 6. Each of these peaks presents seasonality in the time series. The most dominant frequency is 5. Since the data is monthly, this translates to 5 cycles per year. The period is 2.4 months which means that the dominant cycle repeats itself every 2.4 months. This may suggest that the data has a short-term periodic pattern that repeats quite frequently. The second dominant frequency is 3. This means that there are 3 cycles per year and it repeats every 4 months. These two dominant frequencies suggest that there are at least 2 potentially significant cycles in the data. The cause could probably be due to seasonality, perhaps people often travel outside of the U.S. for the holidays and come back through the border. It could also be that students tend to return home for break and return every 3-4 months. There are many possible explanations.

From this smoother periodogram, it is much easier to see the spectrum peaks at frequency 3 and 5, as well as making the smaller cycles at 2, 4, and possibly 6 more apparent.

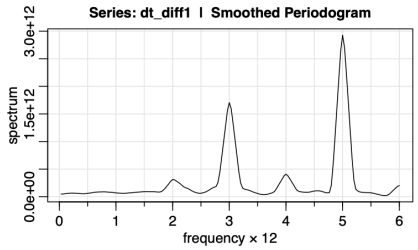


Figure 7: This is the smoothed periodogram of the time series through the Modified Daniell Kernel.

For significance testing of the frequencies, the null hypothesis is that the observed spectral power at a particular frequency is due to random noise, in that it is not a true periodic signal. The alternative is that the spectral power at a particular frequency is a true periodic signal. We reject the null hypothesis if a spectrum value is significantly higher than the upper bound of the confidence interval. For frequencies, since the spectrum is much lower than the upper bound, we fail to reject the null hypothesis. We don't have enough evidence to show that the frequencies are truly a periodic signal.

Table 2: Top Spectral Frequencies with Confidence Intervals

Frequency	Spectrum	Lower Bound	Power	Upper Bound
5	1.789447e+13	2.476765e+12	1.789447e+13	4.127352e+14
3	8.888886e+12	1.230306e+12	8.888886e+12	2.050217e+14
4	2.301665e+12	3.185724e+11	2.301665e+12	5.308780e+13
6	1.554545e+12	2.151638e+11	1.554545e+12	3.585551e+13
2	1.147978e+12	1.588910e+11	1.147978e+12	2.647805e+13

Conclusion and Future Studies

The purpose of this project was to fit a SARIMA model to the time series data for the number of people entering the U.S. through all 28 ports of entry along the U.S.-Mexico Border from January 1996 to December 2024. The data was collected by the U.S. Customs and Border Patrol, thus the data didn't account for any illegal entries. It can be inferred that after the U.S. passed the Immigration Act of 1990 which reduced restrictions on immigration, the number of people entering the U.S. through the U.S. Mexico began to rise and peaked in 1999 to 2000. The numbers began to fall until they hit a low point in 2011 probably due to the 9/11 attack. Another even lower point than ever according to the COVID-19 pandemic where the U.S. strictly limited people from traveling to stop the spread of the disease. After the pandemic, the number of entries seems like it's correcting itself to match the pattern before COVID-19, according to the forecast. The forecast was made using the SARIMA(2, 1, 0)(2, 0, 0)[12] model, which was developed via the auto_arima() function in R. The fit is relatively good with a low AIC of about 30.90. The model is good at predicting the middle lag times of the time series but falls off in performance for the first few lags and higher lag times. The model perhaps failed to capture seasonality in data. After conducting spectral analysis, a total of 2 dominant peaks were found at frequency 5 and 3. Smaller peaks were found at frequencies of 2, 4, and 6. Confidence intervals were made for each peak to determine if they were significant. It was found that none of the peaks were statistically significant implying no true periodic signal.

There were about 3 outlier values in the data set. There only was one which was for sure due to COVID-19. The other two were not looked into. With COVID-19 impacting the time series so greatly, perhaps another analysis of the same time series should be conducted to examine the data before and after COVID. However, the problem with this is that as of now in 2025, This is only 5 years after COVID and there might not be enough data yet. Perhaps in a decade or so, a future analysis should return to this project and conduct the suggested analysis.
For what it is the data in time series only accounted for 28 years of people entering the U.S through the U.S. Mexico Border. There data since the beginning of the border's establishment since 1909 that was not included. It might be worthwhile to include those earlier data in the

analysis to see how the U.S. immigration laws in the past has truly shown itself in the time series.

References

Border Crossing Entry Data. (2025). [Dataset]. Bureau of Transportation Statistics. https://catalog.data.gov/dataset/border-crossing-entry-data-683ae

Gramlich, J. (2024, October 1). Migrant encounters at U.S.-Mexico border have fallen sharply in 2024. *Pew Research Center*. https://www.pewresearch.org/short-reads/2024/10/01/migrant-encounters-at-u-s-mexico-border-have-fallen-sharply-in-2024/

Passel, M. M. and J. S. (2024, September 27). What the data says about immigrants in the U.S. *Pew Research Center*. https://www.pewresearch.org/short-reads/2024/09/27/key-findings-about-us-immigrants/

Appendix

The Augmented Dickey-Fuller Test was also used to see if the differenced time series is stationary.

```
    Augmented Dickey-Fuller Test

data:  dt_diff1
Dickey-Fuller = -6.7326, Lag order = 7, p-value = 0.01
alternative hypothesis: stationary
```

The p-value of 0.01 suggest that at the 0.05 significance level, there is enough evidence to show that the time series stationary after differencing.

```{r, message = FALSE, echo = FALSE}
library(tidyverse)
library(janitor)
library(ggplot2)
library(tseries)
library(astsa)
library(forecast)
```

```{r, echo = FALSE}
df <- read_csv("./data/Border_Crossing_Entry_Data.csv", show_col_types = FALSE) %>%
clean_names()
df <- df %>% filter(border == "US-Mexico Border")
```

```{r, echo = FALSE}
# Function to convert "Month Year" format into index numbers
convert_to_index <- function(month_year) {
  # Define the list of months
  months <- c("Jan", "Feb", "Mar", "Apr", "May", "Jun", "Jul", "Aug", "Sep", "Oct", "Nov", "Dec")

  # Extract the month and year from the input
  month <- substr(month_year, 1, 3)  # Extract first 3 letters (month)
  year <- as.integer(substr(month_year, 5, 8))  # Extract year

  # Find the month index (1 to 12)
  month_index <- match(month, months)

  # Calculate the index number based on year and month
  index <- (year - min(year)) * 12 + month_index  # Adjust for starting year
```

```
  return(index)
}
```

```{r, echo = FALSE}
df1 <- df
date_num <- convert_to_index(df$date)

df1 <- df1 %>% mutate(
  date_num = date_num
)

## group all entries and sum all entries across the US border.
dt <- df1 %>% group_by(date) %>% summarise(total_entry = sum(value), .groups = "drop")

dt_preserved <- dt %>% mutate(
  month = factor(str_extract(date, "^[A-Za-z]{3}"), levels = month.abb),
  year = as.numeric(str_extract(date, "\\d{4}"))) %>% arrange(year, month)
# dt_preserved %>% select(total_entry)
```

```{r, echo = FALSE}
dt_preserved <- dt_preserved %>%
  mutate(
    month = factor(str_extract(date, "^[A-Za-z]{3}"), levels = month.abb),
    year = as.numeric(str_extract(date, "\\d{4}")),
    # Create a proper Date column by combining year and month
    date_formatted = as.Date(paste(year, match(month, month.abb), "01", sep = "-"))
  ) %>%
  arrange(year, month)
```

```{r, echo = FALSE}
# ordered container of total entries
dt <- dt_preserved %>% select(total_entry)
# turns dt into a time series object
dt <- dt %>% ts(frequency = 12)
```
```{r, echo = FALSE, fig.width=5, fig.height=3, fig.cap = "This plot shows the number of people entering the U.S through the U.S. Mexico border entry from 1996 to 2024 under 4 different presidencies, shown through 4 different colors."}
graph_df <- dt_preserved %>%
  group_by(year) %>%
```

```
  summarize(all_entry = sum(total_entry), .groups = "drop")
# Define the U.S. Presidents and their terms, adjusted for the data starting from 1996
presidents <- data.frame(
  president = c("Bill Clinton", "George W. Bush", "Barack Obama", "Donald Trump", "Joe Biden"),
  start_year = c(1993, 2001, 2009, 2017, 2021),  # Adjusted start year for Bill Clinton
  end_year = c(2001, 2009, 2017, 2021, 2025)  # End year for Joe Biden can be adjusted when necessary
)

# Create a simple line plot using base R
plot(graph_df$year, graph_df$all_entry, type = "l", col = "black", lwd = 2,
     xlab = "Year", ylab = "Total Entries", main = "U.S. Mexico Border Entries under each President")

# Add points to the plot
points(graph_df$year, graph_df$all_entry, col = "red", pch = 19)

# Add shaded zones for each president's term using `rect()`
for (i in 1:nrow(presidents)) {
  rect(presidents$start_year[i], par("usr")[3],
       presidents$end_year[i], par("usr")[4],
       col = adjustcolor(c("blue", "orange", "green", "red", "purple")[i], alpha.f = .3), border = NA)
}

# Add a legend to describe the presidents
legend("topright", legend = presidents$president, fill = c("blue", "orange", "green", "red", "purple"),
       title = "Presidents", cex = 0.5)
```

```{r, echo = FALSE,  fig.cap= "The time series plot below is the number of U.S entry through the U.S. Mexico border time series broken down into monthly data over the past 28 years. The time series starts at time 0, which means on January 1996. The data ends on year 28 which is December 2024."}
# fig.width=5, fig.height=3,
plot.ts(dt, main = "Monthly U.S. Mexico Border Entry Trend", ylab = "Number of Entry")
```

```{r, echo = FALSE,  fig.cap = "The time series becomes stationary after taking the first difference."}
# fig.width=6, fig.height=4
dt_diff1 <- diff(dt)
plot.ts(dt_diff1, main = "Differenced Border Crossing Time Series", xlab = "Time", ylab = "Total Border Entry")
```

```{r, echo = FALSE}
adf.test(dt_diff1) # differenced time series is Stationary
```

```{r, echo = FALSE, fig.width=5, fig.height=3, fig.cap = "These two figures show the ACF and PACF Plot of the Entry through U.S. Mexico Border Time Series"}
```

```
par(mfrow = c(1, 2))
dt_diff1 <- diff(dt)
dt_acf <- acf(dt_diff1, main = "ACF Plot")
dt_pacf <- pacf(dt_diff1, main = "PACF Plot")
```
```{r, echo = FALSE}
model <- auto.arima(dt)
sarima21020012 <- sarima(dt, p = 2, d = 1, q = 0, P = 2, D =0, Q = 0, S = 12, detail = FALSE)$table
```
```{r, echo = FALSE,  fig.width=5, fig.height=3, fig.cap = "This chart shows the forecasted total number of people entering the U.S. through all 28 ports of entry on the U.S. Mexico border for the next 12 months after December 2024."}
sarima_forecast <- sarima.for(dt, n.ahead = 12, p = 2, d = 1, q = 1, D = 1, Q = 1, S = 12, main = "Forecast of U.S. Entries for Next 12 Months")
```
```{r, echo = FALSE,  fig.width=5, fig.height=3}
library(kableExtra)
library(scales)
# Data as a data frame
monthly_data <- data.frame(
  Month = c("Jan", "Feb", "Mar", "Apr", "May", "Jun", "Jul", "Aug", "Sep", "Oct", "Nov", "Dec"),
  Value = c(23874170, 22253802, 24603752, 23217083, 23913457, 23224881,
            24319292, 24205166, 23052599, 24081420, 23938205, 25447911) %>% comma()
)

# Create a nice-looking table
kable(monthly_data, format = "latex", booktabs = TRUE, caption = "Predicted Number of Entries through U.S. Mexico Border for Next 12 Months") %>%
  kable_styling(latex_options = c("striped", "hold_position", "scale_down"))
```
```{r, echo = FALSE,  fig.width=5, fig.height=3}
border_spec <- mvspec(dt_diff1, log = "n")
```
```{r, echo = FALSE, fig.cap = "This is the smoothed periodogram of the time series through the Modified Daniell Kernel.", message = FALSE, fig.width=5, fig.height=3}
border_per_tapeSpan <- mvspec(dt_diff1, taper = 0.1, span = c(7, 7))
```
```{r, echo = FALSE}
top_indices <- order(border_spec$spec, decreasing = TRUE )[1:5]

top_frequencies <- border_spec$freq[top_indices]
ci_l <- NULL
ci_u <- NULL
power <- NULL
```

```
df <- border_spec$df
U <- qchisq(0.025, df)
L <- qchisq(0.975, df)
period <- NULL

for (i in 1:length(top_frequencies)){
  period <- append(period, (1/(top_frequencies[i])))
  ci_l <- append(ci_l, (border_spec$spec[top_indices[i]] / L))
  ci_u <- append(ci_u, (border_spec$spec[top_indices[i]] / U))
  power <- append(power, border_spec$spec[top_indices[i]])
}

top_freq_table <- tibble(
  "Frequency" = top_frequencies,
  "Spectrum" = border_spec$spec[top_indices],
  "Lower Bound" = ci_l,
  "Power" =  power,
  "Upper Bound" = ci_u
) %>% arrange(desc(Spectrum))

top_freq_table %>%
  kbl(caption = "Top Spectral Frequencies with Confidence Intervals") %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed", "responsive"))
```