Telling the 2011-2019 Story of U.S. Arrivals Through I-94

By Karen Wu, Reed Leon-Hinton, Akshay Krishnan

Duke University, Nicholas School of the Environment

Abstract

Data analysis and forecasting are important tools in not only understanding past trends in travel to a country, but also in tourism and immigration policymaking. This study focuses on analyzing entry patterns into the United States from nine world regions in order to understand its major determinants. The study makes use of multivariate time series techniques based on the federal I-94 form during the period 2011-2019 to inform future planning practices. The analysis of existing data suggests that multiple factors affect the travel trends within the United States, ranging from civil wars, USD currency exchange rates, presidential elections, to pandemics. The pattern is also highly dependent on region and is dramatically affected by local incidents. Based on the analysis, past trends can help make forecasts for up to two years into the future with a high level of confidence. Nonetheless, it is important to note that the I-94 exhibits certain limitations such as lack of immigration data and exclusion of certain countries. Lastly, certain incidents such as the COVID-19 outbreak can completely alter future projection, which makes future projections highly susceptible to sharp positive or negative fluctuations.

Introduction

The I-94 Form is a tool utilized by the U.S. Customs and Border Protection which collects a variety of information about foreign individuals arriving or departing from the United States (I-94 Official Website 2020). The form collects demographic information along with the type of visa the individual is traveling with and where they are traveling to or from. It is completed electronically by nearly all visitors to the United States except those that are citizens of Mexico or Canada and traveling over the respective borders. Other notable exceptions are returning resident aliens, aliens with an immigrant visa, and Canadian citizens in transit between countries (I-94 Official Website FAQ 2020).

The I-94 Form presents a unique situation where data is collected in a robust and thorough manner. This allows for the construction of robust and high-quality forecast models, as the data on which the model is being created is clean and descriptive of real patterns. The U.S. National Travel & Tourism Office works with the Department of Homeland Security and Customs and Border Protection to manage the I-94 dataset and make the information available to the general public (NTTO 2020). They collect and provide the data for both country of residence and country of origin, along with the month and year of arrival.

Tourism and visitation to the United States represent a significant economical and societal benefit. Thus, being able to predict the amount of arrivals and the regions which will have the greatest impact in the future would be extremely valuable information. This study attempts to do just that: formulate a time series model which forecasts the arrivals to the United States using the I-94 Form country of residence dataset provided by the National Travel & Tourism Office.

Methods

In order to analyze the seasonality of the dataset, first I created several time series objects using the *ts()* function available in R statistical programming while in R-Studio. I created a time series object for each world region represented on the original dataset and stated a frequency of twelve, denoting the twelve months in a year. Then, I used the *decompose()* function on each of the world region time series objects to separate and store the seasonality coefficients for all the months between January 2011 and February 2020. Positive seasonality coefficients represent a net influx of arrivals into the U.S. for a particular month of a certain year, while negative seasonality coefficients represent a net outflux of "arrivals" into the U.S. for a particular month of a certain year. Then I imported these seasonality coefficients of each world region for January 2011 to February 2020 into Tableau Desktop data visualization software, in order to compare average seasonal coefficients amongst the world regions.

Below is the R code used to extract seasonality coefficients for each world region:

```
Western_Europe<-ts(data$`Western Europe`,frequency=12,start=c(2011,1))
Asia<-ts(data$Asia,frequency=12,start=c(2011,1))
South_America<-ts(data$`South America`,frequency=12,start=c(2011,1))
Caribbean<-ts(data$Caribbean,frequency=12,start=c(2011,1))
Oceana<-ts(data$Oceana,frequency=12,start=c(2011,1))
Central_America<-ts(data$`Central America`,frequency=12,start=c(2011,1))
Middle_East<-ts(data$`Middle East`,frequency=12,start=c(2011,1))
Eastern_Europe<-ts(data$`Eastern Europe`,frequency=12,start=c(2011,1))
Africa<-ts(data$Africa,frequency=12,start=c(2011,1))

####

decomp_WestEurope<-decompose(Western_Europe)
write.csv(decomp_WestEurope$seasonal,"~/Duke/TimeSeriesProject/WestEurope.csv")

decomp_Asia<-decompose(Asia)
write.csv(decomp_Asia$seasonal,"~/Duke/TimeSeriesProject/Asia.csv")

decomp_SouthAmerica<-decompose(South_America)
write.csv(decomp_SouthAmerica$seasonal,"~/Duke/TimeSeriesProject/SouthAmerica.csv")

decomp_Caribbean<-decompose(Caribbean)
write.csv(decomp_Caribbean$seasonal,"~/Duke/TimeSeriesProject/Caribbean.csv")

decomp_Oceana<-decompose(Oceana)
write.csv(decomp_Oceana$seasonal,"~/Duke/TimeSeriesProject/Oceana.csv")
```

```
decomp_CentralAmerica<-decompose(Central_America)
write.csv(decomp_CentralAmerica$seasonal,"~/Duke/TimeSeriesProject/CentralAmerica.csv")

decomp_MiddleEast<-decompose(Middle_East)
write.csv(decomp_MiddleEast$seasonal,"~/Duke/TimeSeriesProject/MiddleEast.csv")

decomp_EasternEurope<-decompose(Eastern_Europe)
write.csv(decomp_EasternEurope$seasonal,"~/Duke/TimeSeriesProject/EastEurope.csv")

decomp_Africa<-decompose(Africa)
write.csv(decomp_Africa$seasonal,"~/Duke/TimeSeriesProject/Africa.csv")
```

Besides seasonality, other crucial components when conducting any sort of time series analysis is an understanding of the dataset's original observed values, trends, and the presence of randomness. Each of these facets contribute to the accuracy and precision of forecast models created and, in some cases, require removal from the time series or differencing to eliminate in order to produce a quality prediction. The first step in analyzing the general patterns within the time series is graphing the observed values to identify any potential anomalous values, obvious seasonality, or distinguishable trends. Each of the different regions observed values were graphed in a matrix along with simple linear regression trend lines (green) and the mean value throughout the period (yellow) in Figure 1 below. The R code utilized to generate these graphs is included below:

```
#Loading the data
 regions=read.csv(file="World_Regions.csv",header=TRUE)
regions=regions[1:110,]
rownames(regions)=regions[,2]
regions=regions[,3:11]

#Creating a dates dataset
regions_month=read.csv(file="Dates.csv",header=TRUE)
regions_month=regions_month[1:110,3]
dates=as.Date(regions_month,"%m/%d/%Y")
print(dates)

#Creating the time series
tsregions=ts(data=regions,frequency=12)

#plotting the lines

par(mfrow=c(3,3))

plot(dates,tsregions[,1],type="l", xlab="Month (Jan 2011 - Feb 2020)",
```

```
    ylab="Arrivals (# people)",main="Western Europe")
abline(h=mean(tsregions[,1]),col="orange")
abline(lm(tsregions[,1]~dates),col="forestgreen")

plot(dates,tsregions[,2],type="l", xlab="Month (Jan 2011 - Feb 2020)",
    ylab="Arrivals (# people)",main="Asia")
abline(h=mean(tsregions[,2]),col="orange")
abline(lm(tsregions[,2]~dates),col="forestgreen")

plot(dates,tsregions[,3],type="l", xlab="Month (Jan 2011 - Feb 2020)",
    ylab="Arrivals (# people)",main="South America")
abline(h=mean(tsregions[,3]),col="orange")
abline(lm(tsregions[,3]~dates),col="forestgreen")

plot(dates,tsregions[,4],type="l", xlab="Month (Jan 2011 - Feb 2020)",
    ylab="Arrivals (# people)",main="Caribbean")
abline(h=mean(tsregions[,4]),col="orange")
abline(lm(tsregions[,4]~dates),col="forestgreen")

plot(dates,tsregions[,5],type="l", xlab="Month (Jan 2011 - Feb 2020)",
    ylab="Arrivals (# people)",main="Oceana")
abline(h=mean(tsregions[,5]),col="orange")
abline(lm(tsregions[,5]~dates),col="forestgreen")

plot(dates,tsregions[,6],type="l", xlab="Month (Jan 2011 - Feb 2020)",
    ylab="Arrivals (# people)",main="Central America")
abline(h=mean(tsregions[,6]),col="orange")
abline(lm(tsregions[,6]~dates),col="forestgreen")

plot(dates,tsregions[,7],type="l", xlab="Month (Jan 2011 - Feb 2020)",
    ylab="Arrivals (# people)",main="Middle East")
abline(h=mean(tsregions[,7]),col="orange")
abline(lm(tsregions[,7]~dates),col="forestgreen")

plot(dates,tsregions[,8],type="l", xlab="Month (Jan 2011 - Feb 2020)",
    ylab="Arrivals (# people)",main="Eastern Europe")
abline(h=mean(tsregions[,8]),col="orange")
abline(lm(tsregions[,8]~dates),col="forestgreen")

plot(dates,tsregions[,9],type="l", xlab="Month (Jan 2011 - Feb 2020)",
    ylab="Arrivals (# people)",main="Africa")
abline(h=mean(tsregions[,9]),col="orange")
abline(lm(tsregions[,9]~dates),col="forestgreen")
```

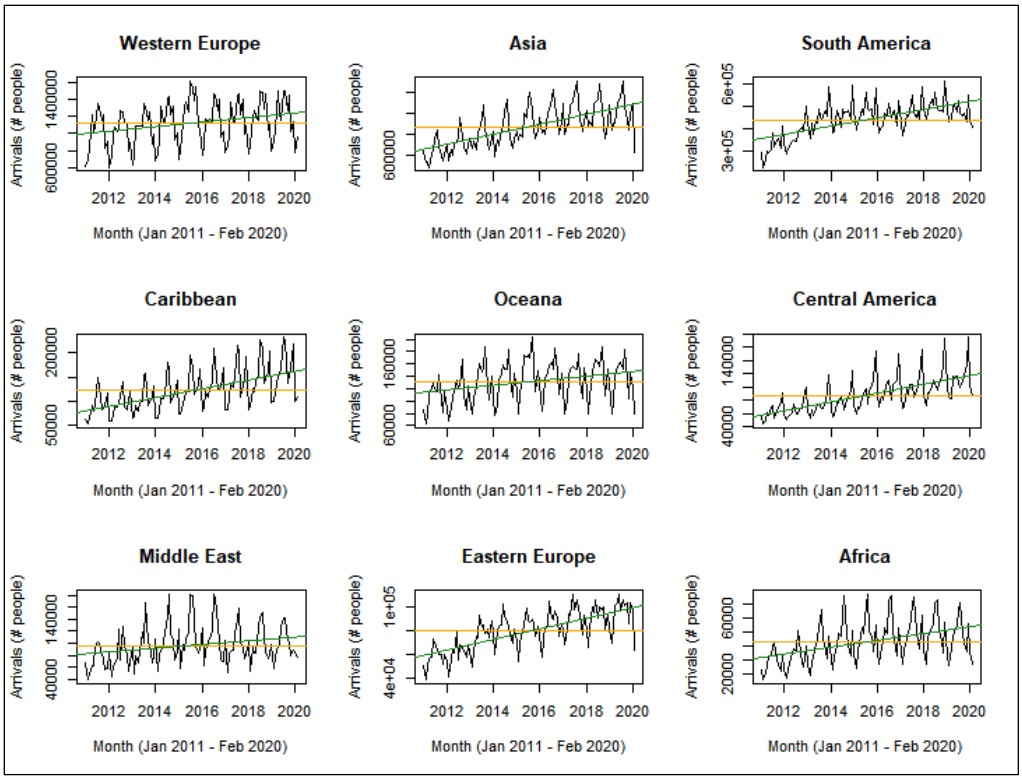

Figure 1: Matrix of the observed value for each region with linear regression trend lines in green and the mean value for the entire time period in yellow.

After looking at the original, observed time series the seasonality was removed to better analyze the trend in the deseasoned data. This transformation was conducted by running the decompose function and removing the seasonal component identified from the original time series using the *seasadj* function. The results of the decompose function also provided insight into the trend component of each region's time series along with the random behavior, as seen in Figure 2. Additionally, the deseasoned data was graphed in a matrix with the same order as the matrix of the original series to allow easy comparison in Figure 3. The R code used to create both of these graphs is included below:

```
#Making individual time series for each dataset

ts_WE=ts(data=tsregions[,1],frequency=12) #Western Europe
ts_AS=ts(data=tsregions[,2],frequency=12) #Asia
ts_SA=ts(data=tsregions[,3],frequency=12) #South America
ts_CB=ts(data=tsregions[,4],frequency=12) #Caribbean
ts_OC=ts(data=tsregions[,5],frequency=12) #Oceana
ts_CA=ts(data=tsregions[,6],frequency=12) #Central America
ts_ME=ts(data=tsregions[,7],frequency=12) #Middle East
ts_EE=ts(data=tsregions[,8],frequency=12) #Eastern Europe
ts_AF=ts(data=tsregions[,9],frequency=12) #Africa

#Looking at the decompositions
```

```
decomp_WE=decompose(ts_WE,"additive")
plot(decomp_WE)

decomp_AS=decompose(ts_AS,"additive")
plot(decomp_AS)

#Making a title for the decompose function for the Asian Region
plot(cbind(Observed = decomp_AS$x,Trend=decomp_AS$trend,
        Seasonal=decomp_AS$seasonal,Random=decomp_AS$random),main="Decomposition
of Asian Region Time Series")
plot(decomp_AS)

decomp_SA=decompose(ts_SA,"additive")
plot(decomp_SA)

decomp_CB=decompose(ts_CB,"additive")
plot(decomp_CB)

decomp_OC=decompose(ts_OC,"additive")
plot(decomp_OC)

decomp_CA=decompose(ts_CA,"additive")
plot(decomp_CA)

decomp_ME=decompose(ts_ME,"additive")
plot(decomp_ME)

decomp_EE=decompose(ts_EE,"additive")
plot(decomp_EE)

decomp_AF=decompose(ts_AF,"additive")
plot(decomp_AF)

#Deseasoning the data and graphing them

par(mfrow=c(3,3))

deseason_WE=seasadj(decomp_WE)
plot(dates,deseason_WE,type="l", xlab="Month (Jan 2011 - Feb 2020)",
    ylab="Arrivals (# people)",main="Western Europe")
abline(lm(deseason_WE~dates),col="forestgreen")

deseason_AS=seasadj(decomp_AS)
plot(dates,deseason_AS,type="l", xlab="Month (Jan 2011 - Feb 2020)",
```

```
    ylab="Arrivals (# people)",main="Asia")
abline(lm(deseason_AS~dates),col="forestgreen")

deseason_SA=seasadj(decomp_SA)
plot(dates,deseason_SA,type="l", xlab="Month (Jan 2011 - Feb 2020)",
    ylab="Arrivals (# people)",main="South America")
abline(lm(deseason_SA~dates),col="forestgreen")

deseason_CB=seasadj(decomp_CB)
plot(dates,deseason_CB,type="l", xlab="Month (Jan 2011 - Feb 2020)",
    ylab="Arrivals (# people)",main="Caribbean")
abline(lm(deseason_CB~dates),col="forestgreen")

deseason_OC=seasadj(decomp_OC)
plot(dates,deseason_OC,type="l", xlab="Month (Jan 2011 - Feb 2020)",
    ylab="Arrivals (# people)",main="Oceana")
abline(lm(deseason_OC~dates),col="forestgreen")

deseason_CA=seasadj(decomp_CA)
plot(dates,deseason_CA,type="l", xlab="Month (Jan 2011 - Feb 2020)",
    ylab="Arrivals (# people)",main="Central America")
abline(lm(deseason_CA~dates),col="forestgreen")

deseason_ME=seasadj(decomp_ME)
plot(dates,deseason_ME,type="l", xlab="Month (Jan 2011 - Feb 2020)",
    ylab="Arrivals (# people)",main="Middle East")
abline(lm(deseason_ME~dates),col="forestgreen")

deseason_EE=seasadj(decomp_EE)
plot(dates,deseason_EE,type="l", xlab="Month (Jan 2011 - Feb 2020)",
    ylab="Arrivals (# people)",main="Eastern Europe")
abline(lm(deseason_EE~dates),col="forestgreen")

deseason_AF=seasadj(decomp_AF)
plot(dates,deseason_AF,type="l", xlab="Month (Jan 2011 - Feb 2020)",
    ylab="Arrivals (# people)",main="Africa")
abline(lm(deseason_AF~dates),col="forestgreen")
```

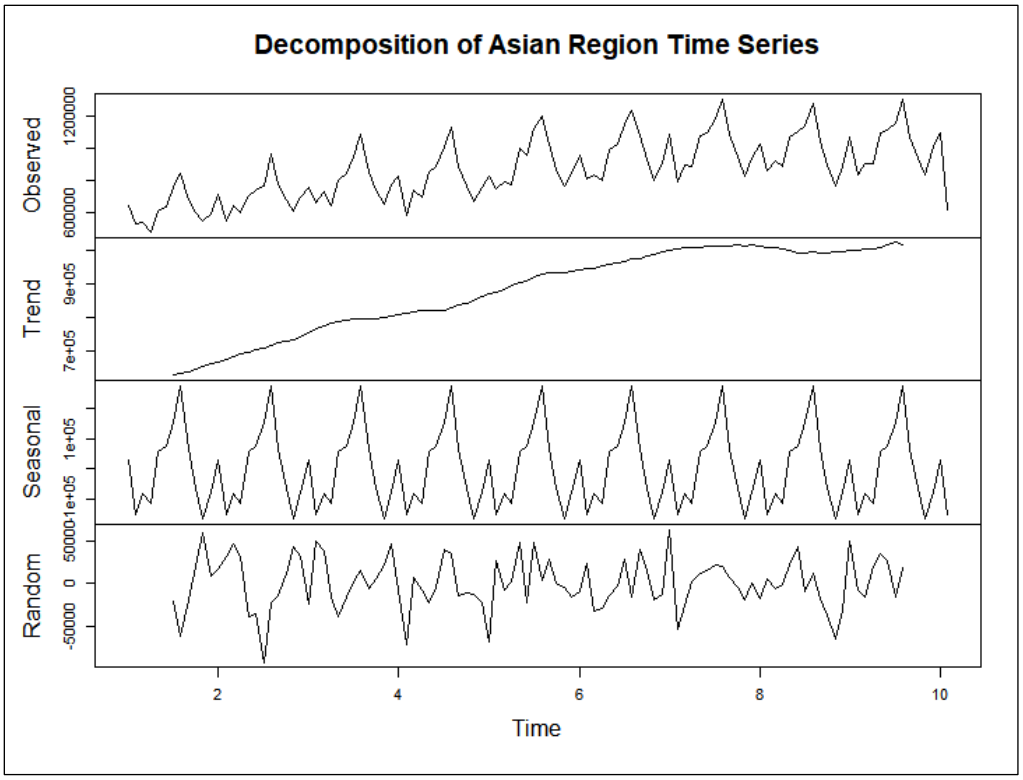

Figure 2: Results of the decompose function on the "Asia" Region arrivals to the U.S. broken up into the observed data, the trend component, the seasonal trend, and the random variations in the dataset.

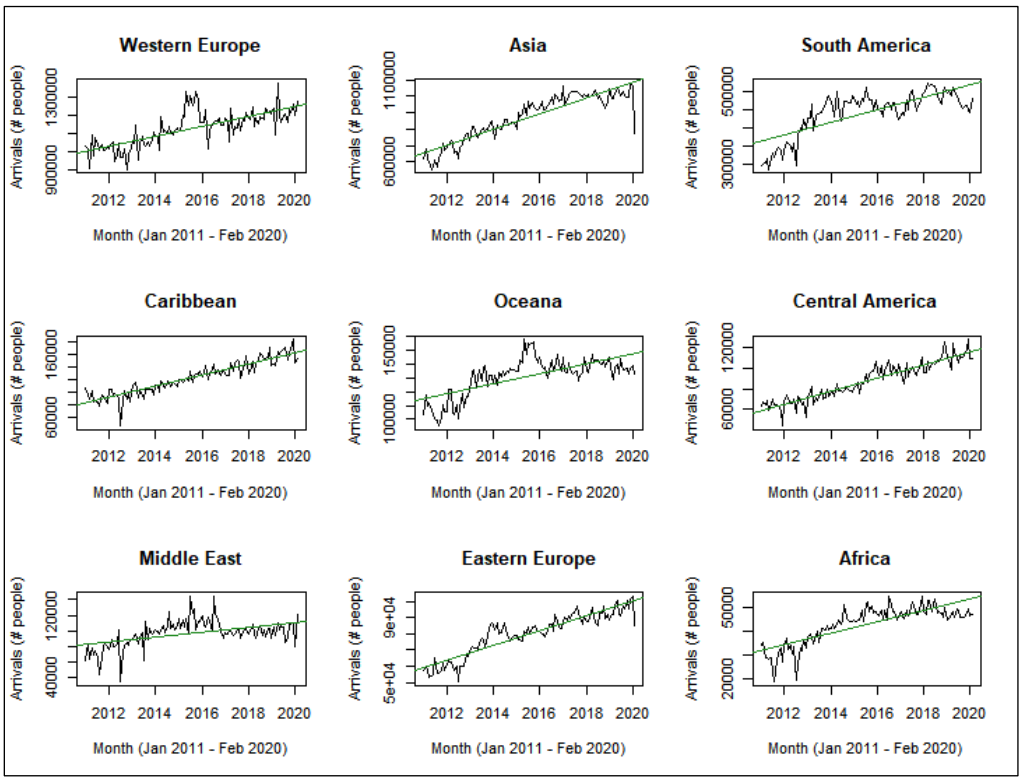

Figure 3: Matrix of the deseasoned time series for each region with linear regression trend lines in green.

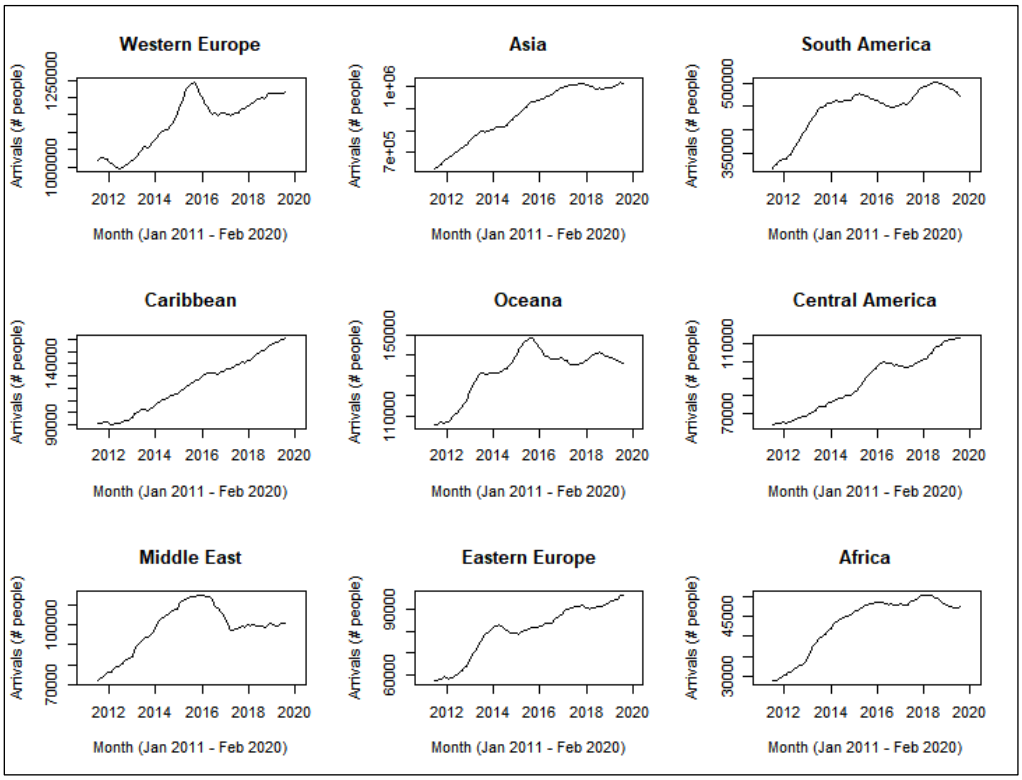

Figure 4: Matrix of the trend component calculated as part of the decompose function for each region.

After creating the matrix of deseasoned time series, the random component was still displayed on the graphs, which can make the interpretation of the trend slightly more challenging. Therefore, the trend components from the decompose functions were isolated and plotted in another graph matrix to verify the pattern in the deseasoned time series (Figure 4). The trend components from the decompose function supported the decrease seen in Western Europe and Oceana, while also revealing similar trends in many other regions: Middle East, Central America, and South America. This suggests that there may be an external factor, not captured by the model, affecting the number of arrivals during that time period. The first factor considered was the value of the U.S. Dollar, as a low value would likely increase tourism as visitors have a favorable exchange rate from other currencies to the USD. The trend from Western Europe was compared to the trend in USD values gathered from MacroTrends to assess potential correlation in Figure 5. The R code utilized to graph Figures 4 and 5 is included below Figure 5:

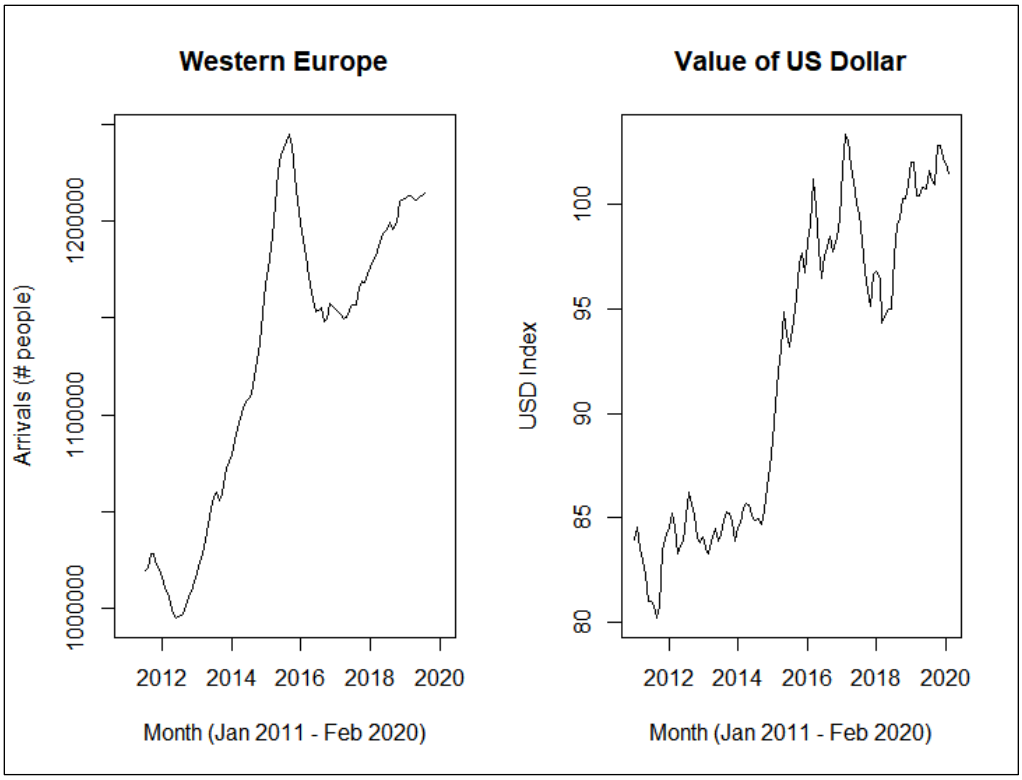

Figure 5: Comparison between the increase in arrivals from Western Europe and the corresponding value of the U.S. Dollar ($).

```
#Comparing the trends in the data

par(mfrow=c(3,3))

plot(dates,decomp_WE$trend,type="l", xlab="Month (Jan 2011 - Feb 2020)",
    ylab="Arrivals (# people)",main="Western Europe")

plot(dates,decomp_AS$trend,type="l", xlab="Month (Jan 2011 - Feb 2020)",
    ylab="Arrivals (# people)",main="Asia")

plot(dates,decomp_SA$trend,type="l", xlab="Month (Jan 2011 - Feb 2020)",
    ylab="Arrivals (# people)",main="South America")

plot(dates,decomp_CB$trend,type="l", xlab="Month (Jan 2011 - Feb 2020)",
    ylab="Arrivals (# people)",main="Caribbean")

plot(dates,decomp_OC$trend,type="l", xlab="Month (Jan 2011 - Feb 2020)",
    ylab="Arrivals (# people)",main="Oceana")

plot(dates,decomp_CA$trend,type="l", xlab="Month (Jan 2011 - Feb 2020)",
    ylab="Arrivals (# people)",main="Central America")
```

```
plot(dates,decomp_ME$trend,type="l", xlab="Month (Jan 2011 - Feb 2020)",
    ylab="Arrivals (# people)",main="Middle East")

plot(dates,decomp_EE$trend,type="l", xlab="Month (Jan 2011 - Feb 2020)",
    ylab="Arrivals (# people)",main="Eastern Europe")

plot(dates,decomp_AF$trend,type="l", xlab="Month (Jan 2011 - Feb 2020)",
    ylab="Arrivals (# people)",main="Africa")

#graph of USD value
USD_value=read.csv(file="USD_Historical.csv",header=TRUE)
USD_valuets=ts(USD_value[,2],frequency = 12)

par(mfrow=c(1,2))

plot(dates,decomp_WE$trend,type="l", xlab="Month (Jan 2011 - Feb 2020)",
    ylab="Arrivals (# people)",main="Western Europe")
plot(dates,USD_valuets[11:120],type="l", xlab="Month (Jan 2011 - Feb 2020)",
    ylab="USD Index",main="Value of US Dollar")
```

The next step in the process was to fit a model to illustrate the trend in all these datasets i.e. in each of the nine regions. This is important not only to analyze the data, but also to test the feasibility of the model for future projections and forecasts. The first step in this process was to compare the ACFs and PACFs to determine the appropriate fit. Figure 6 and 7 show the ACFs and PACFs for deseasoned datasets from all 9 regions respectively. The R code for generating these plots is given below:

```
#Comparing ACFs for all regions

par(mfrow=c(3,3))

Acf(deseason_WE,lag.max=40)

Acf(deseason_AS,lag.max=40)

Acf(deseason_SA,lag.max=40)

Acf(deseason_CB,lag.max=40)

Acf(deseason_OC,lag.max=40)

Acf(deseason_CA,lag.max=40)

Acf(deseason_ME,lag.max=40)
```

```
Acf(deseason_EE,lag.max=40)

Acf(deseason_AF,lag.max=40)


#Comparing PACFs for all regions

par(mfrow=c(3,3))

Acf(deseason_WE,lag.max=40,type = "partial")

Acf(deseason_AS,lag.max=40,type = "partial")

Acf(deseason_SA,lag.max=40,type = "partial")

Acf(deseason_CB,lag.max=40,type = "partial")

Acf(deseason_OC,lag.max=40,type = "partial")

Acf(deseason_CA,lag.max=40,type = "partial")

Acf(deseason_ME,lag.max=40,type = "partial")

Acf(deseason_EE,lag.max=40,type = "partial")

Acf(deseason_AF,lag.max=40,type = "partial")
```

The ACF and PACF plots help us make important inferences for fitting the model. Looking at the ACF, most of the plots show a slow decline. While looking at the PACF, we see that a majority of the plots show a clear cutoff after one or two lags. Considering both these trends, we can say that by first impression of the plot, we can infer an AR trend. It is also important to consider that in the ACF plot, majority of the lines are above the significance level. This shows that we may also need to difference the data in order to make it stationary. Looking at the PACF, the lags look to be 1 or 2 order, which indicates that most of the data sets can be made stationary with one or two differences. Lastly, an important takeaway from the plots is that even though the plots show an AR trend at first glance, differencing to remove the unit root may change the fit of the model.

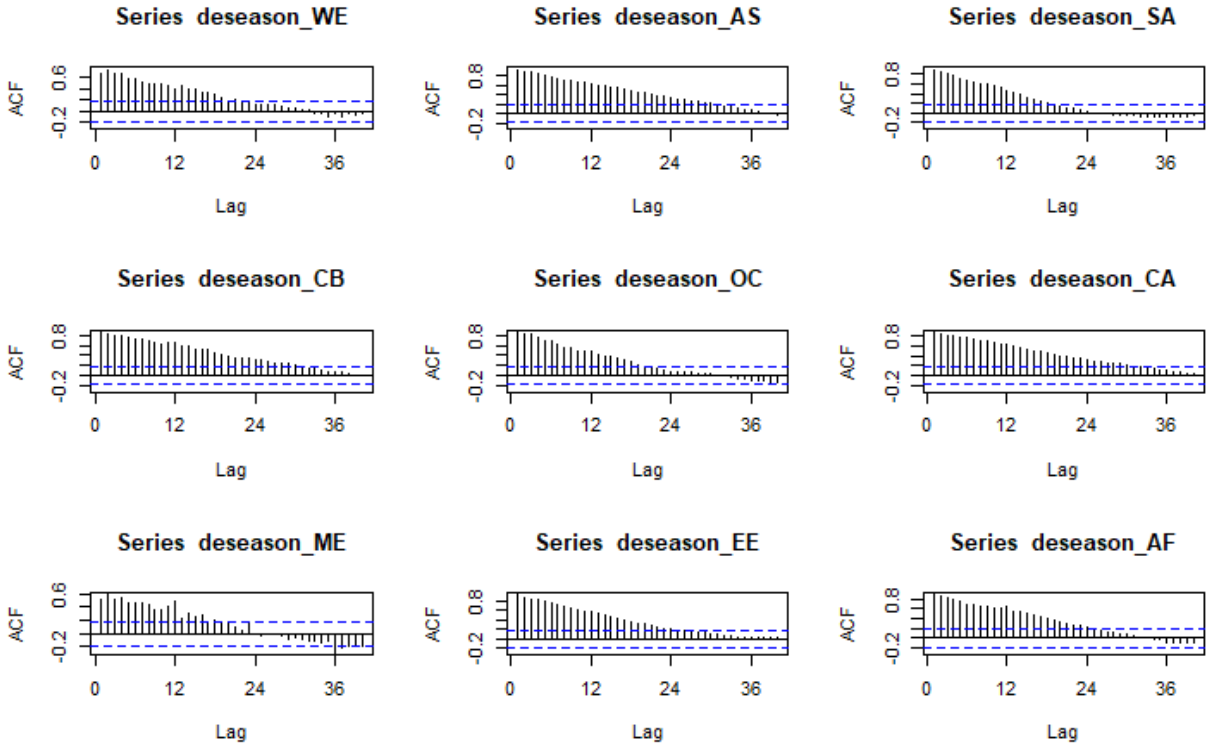

Figure 6: Comparison of ACF plots for deseasoned data from 9 world regions.

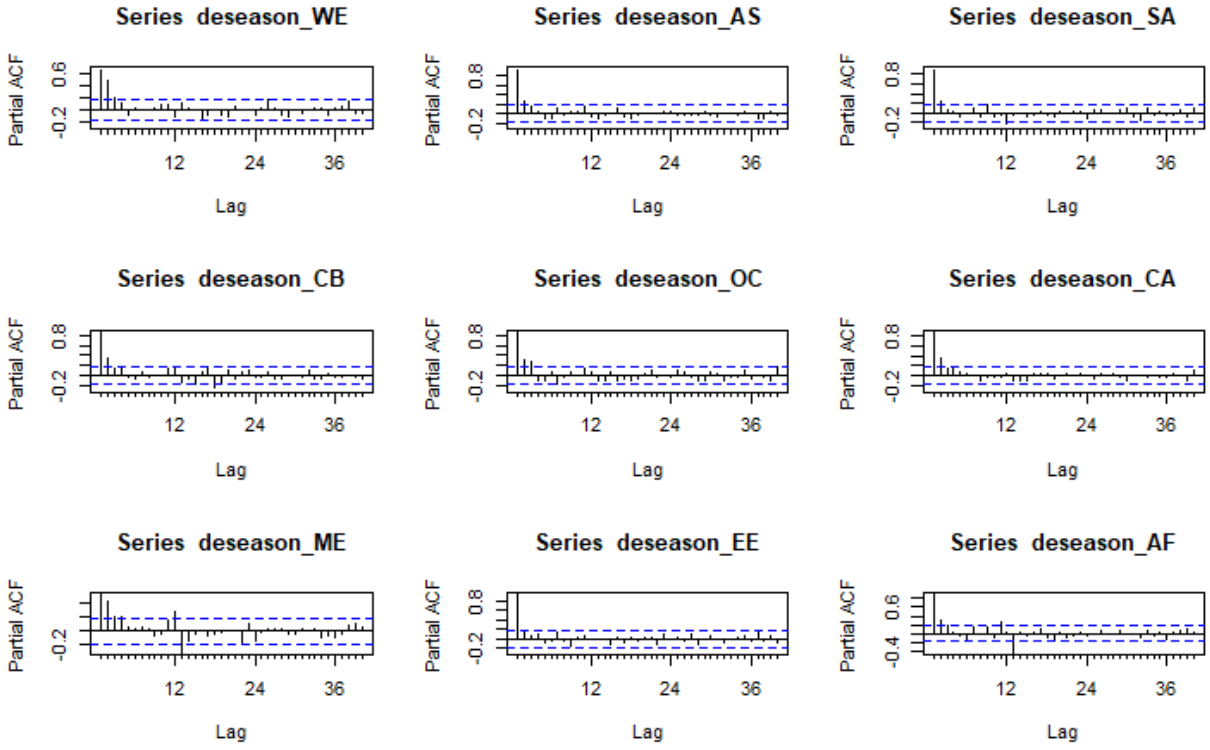

Figure 7: Comparison of PACF plots for seasoned data from 9 world regions.

Model Fitting

Two important characteristics for a dataset to be fitted to an ARIMA model for further analysis are that:

1. The dataset should exhibit a monotonic trend. This means that the variable should consistently increase or decrease through time. This does not necessarily mean that the trend be linear.
2. The dataset should be statistically stationary. This means that the statistical properties of the dataset such as mean, variance, autocorrelation etc. should remain constant over time. It is important for a dataset to be stationary for it to be a useful descriptor of the future behavior.

In order to analyze the monotonicity and stationarity of the dataset, we have to conduct two tests, namely Mann-Kendall test and Augmented Dickey-Fuller (ADF) tests respectively.

Mann-Kendall test

Mann-Kendall test is conducted to analyze if the dataset shows a monotonic trend. The null hypothesis in Mann-Kendall test is that there is no monotonic trend. Thus, we need a small p-value in order to determine that the dataset has a monotonic trend. On conducting the test for all

nine regions, we found that all regions returned a very small p-value. Thus, all the regions have a monotonic trend and are good to be tested for stationarity.

ADF test

The ADF test is conducted to analyze stationarity in the dataset. The null hypothesis in ADF test is that the data exhibits no stationary trend. Thus, we need a small p-value to reject the null. In the nine regions, only three i.e. Western Europe, Caribbean and Central America exhibited a low p-value. Thus, only these three datasets were ideal for fitting to an ARIMA model. For the other datasets, it was important to difference the data in order to achieve stationarity before fitting a model. All the six remaining regions exhibited a low p-value i.e. stationarity after a differencing at lag one.

Fitting a Model

After obtaining the new datasets, the next step was to fit a model. We tried fitting a model both by using the original order, which was obtained from the ACF and PACF for the dataset, and using an Auto.Arima function. In every case, we received a better residual fit for the auto.arima model. It Is also key to note that another important value i.e. AIC should also be checked and we had to make sure that we noticed considerably low AIC values which was a good indication for the fit.

The code used for testing the monotonicity, stationarity and fitting the model is as given below. This also includes plotting the residuals along with their ACF and PACF.

```
#Western Europe Fitting

MannKendall(deseason_WE) #Data has a monotonic trend

adf.test(deseason_WE,alternative="stationary") #Data is stationary

deseason_WE_arima_fit=auto.arima(deseason_WE,max.D=0,max.P = 0,max.Q=0)

print(deseason_WE_arima_fit) # We get order (0,1,2) MA

par(mfrow=c(1,1))

plot(dates[1:110],deseason_WE_arima_fit$residuals,type="l", xlab="Month (Jan 2011 - Feb 2020)",
    ylab="Arrivals (# people)",main="Western Europe Residual")

abline(h=0,col="cadetblue2",lwd=2)
abline(h=mean(deseason_WE_arima_fit$residuals),col="black",lty="dashed",lwd=1)

par(mfrow=c(1,2))

Acf(deseason_WE_arima_fit$residuals,lag.max=40,main = "Western Europe ACF")

Pacf(deseason_WE_arima_fit$residuals,lag.max=40, main = "Western Europe PACF")
```

```
#Asia Fitting

MannKendall(deseason_AS) #Data has a monotonic trend

adf.test(deseason_AS,alternative="stationary") #Data has a unit root

deseason_AS_diff = diff(deseason_AS,differences=1)

adf.test(deseason_AS_diff,alternative="stationary") #Data is now stationary

deseason_AS_diff_arima_fit=auto.arima(deseason_AS_diff,max.D=0,max.P = 0,max.Q=0)

print(deseason_AS_diff_arima_fit) # We get order (0,0,2) MA

par(mfrow=c(1,1))

plot(dates[2:110],deseason_AS_diff_arima_fit$residuals,type="l", xlab="Month (Jan 2011 - Feb 2020)",
    ylab="Arrivals (# people)",main="Asia Residual")

abline(h=0,col="red",lwd=2)
abline(h=mean(deseason_AS_diff_arima_fit$residuals),col="black",lty="dashed",lwd=1)

par(mfrow=c(1,2))

Acf(deseason_AS_diff_arima_fit$residuals,lag.max=40,main="Asia ACF")

Pacf(deseason_AS_diff_arima_fit$residuals,lag.max=40,main="Asia PACF")


#South America Fitting

MannKendall(deseason_SA) #Data has a monotonic trend

adf.test(deseason_SA,alternative="stationary") #Data has unit root

deseason_SA_diff = diff(deseason_SA,differences=1)

adf.test(deseason_SA_diff,alternative="stationary")

deseason_SA_diff_arima_fit=auto.arima(deseason_SA_diff,max.D=0,max.P = 0,max.Q=0)

print(deseason_SA_diff_arima_fit) # We get order (0,0,1) MA

par(mfrow=c(1,1))

plot(dates[2:110],deseason_SA_diff_arima_fit$residuals,type="l", xlab="Month (Jan 2011 - Feb 2020)",
    ylab="Arrivals (# people)",main="South America Residual")

abline(h=0,col="chocolate1",lwd=2)
```

```
abline(h=mean(deseason_SA_diff_arima_fit$residuals),col="black",lty="dashed",lwd=1)

par(mfrow=c(1,2))

Acf(deseason_SA_diff_arima_fit$residuals,lag.max=40, main = "South America ACF")

Pacf(deseason_SA_diff_arima_fit$residuals,lag.max=40, main = "South America PACF")


#Caribbean Fitting

MannKendall(deseason_CB) #Data has a monotonic trend

adf.test(deseason_CB,alternative="stationary") #Data is stationary

deseason_CB_arima_fit=auto.arima(deseason_CB,max.D=0,max.P = 0,max.Q=0)

print(deseason_CB_arima_fit) # We get order (0,1,1) MA

par(mfrow=c(1,1))

plot(dates[1:110],deseason_CB_arima_fit$residuals,type="l", xlab="Month (Jan 2011 - Feb
2020)",
    ylab="Arrivals (# people)",main="Caribbean Residual")

abline(h=0,col="aquamarine2",lwd=2)
abline(h=mean(deseason_CB_arima_fit$residuals),col="black",lty="dashed",lwd=1)

par(mfrow=c(1,2))

Acf(deseason_CB_arima_fit$residuals,lag.max=40,main = "Caribbean ACF")

Pacf(deseason_CB_arima_fit$residuals,lag.max=40, main = "Caribbean PACF")


#Oceana Fitting

MannKendall(deseason_OC) #Data has a monotonic trend

adf.test(deseason_OC,alternative="stationary") #Data has a unit root

deseason_OC_diff = diff(deseason_OC,differences=1)

adf.test(deseason_OC_diff,alternative="stationary") #Data is now stationary

deseason_OC_diff_arima_fit=auto.arima(deseason_OC_diff,max.D=0,max.P = 0,max.Q=0)

print(deseason_OC_diff_arima_fit) # We get order (2,0,0) AR

par(mfrow=c(1,1))
```

```
plot(dates[2:110],deseason_OC_diff_arima_fit$residuals,type="l", xlab="Month (Jan 2011 - Feb 2020)",
    ylab="Arrivals (# people)",main="Oceana Residual")

abline(h=0,col="blue2",lwd=2)
abline(h=mean(deseason_OC_diff_arima_fit$residuals),col="black",lty="dashed",lwd=1)


par(mfrow=c(1,2))

Acf(deseason_OC_diff_arima_fit$residuals,lag.max=40,main="Oceana ACF")

Pacf(deseason_OC_diff_arima_fit$residuals,lag.max=40,main="Oceana PACF")


#Central America Fitting

MannKendall(deseason_CA) #Data has a monotonic trend

adf.test(deseason_CA,alternative="stationary") #Data is stationary

deseason_CA_arima_fit=auto.arima(deseason_CA,max.D=0,max.P = 0,max.Q=0)

print(deseason_CA_arima_fit) # We get order (0,1,1) MA

par(mfrow=c(1,1))

plot(dates[1:110],deseason_CA_arima_fit$residuals,type="l", xlab="Month (Jan 2011 - Feb 2020)",
    ylab="Arrivals (# people)",main="Central America Residual")

abline(h=0,col="blueviolet",lwd=2)
abline(h=mean(deseason_CA_diff_arima_fit$residuals),col="black",lty="dashed",lwd=1)

par(mfrow=c(1,2))

Acf(deseason_CA_arima_fit$residuals,lag.max=40,main="Central America ACF")

Pacf(deseason_CA_arima_fit$residuals,lag.max=40,main="Central America PACF")


#Middle East Fitting

MannKendall(deseason_ME) #Data has a monotonic trend

adf.test(deseason_ME,alternative="stationary") #Data is stationary

deseason_ME_diff = diff(deseason_ME,differences=1)

adf.test(deseason_ME_diff,alternative="stationary")
```

```
deseason_ME_diff_arima_fit=auto.arima(deseason_ME_diff,max.D=0,max.P = 0,max.Q=0)

print(deseason_ME_diff_arima_fit) # We get order (0,0,1) MA

par(mfrow=c(1,1))

plot(dates[2:110],deseason_ME_diff_arima_fit$residuals,type="l", xlab="Month (Jan 2011 - Feb 2020)",
    ylab="Arrivals (# people)",main="Middle East Residual")

abline(h=0,col="darkgoldenrod1",lwd=2)
abline(h=mean(deseason_ME_diff_arima_fit$residuals),col="black",lty="dashed",lwd=1)

par(mfrow=c(1,2))

Acf(deseason_ME_diff_arima_fit$residuals,lag.max=40,main="Middle East ACF")

Pacf(deseason_ME_diff_arima_fit$residuals,lag.max=40,main="Middle East PACF")


#Eastern Europe Fitting

MannKendall(deseason_EE) #Data has a monotonic trend

adf.test(deseason_EE,alternative="stationary") #Data has a unit root

deseason_EE_diff = diff(deseason_EE,differences=1)

adf.test(deseason_EE_diff,alternative="stationary")

deseason_EE_diff_arima_fit=auto.arima(deseason_EE_diff,max.D=0,max.P = 0,max.Q=0)

print(deseason_EE_diff_arima_fit) # We get order (1,0,1) ARMA

par(mfrow=c(1,1))

plot(dates[2:110],deseason_EE_diff_arima_fit$residuals,type="l", xlab="Month (Jan 2011 - Feb 2020)",
    ylab="Arrivals (# people)",main="Eastern Europe Residual")

abline(h=0,col="brown1",lwd=2)
abline(h=mean(deseason_EE_diff_arima_fit$residuals),col="black",lty="dashed",lwd=1)

par(mfrow=c(1,2))

Acf(deseason_EE_diff_arima_fit$residuals,lag.max=40,main="Eastern Europe ACF")

Pacf(deseason_EE_diff_arima_fit$residuals,lag.max=40,main="Eastern Europe PACF")

#Africa Fitting
```

```
MannKendall(deseason_AF) #Data has a monotonic trend

adf.test(deseason_AF,alternative="stationary") #Data has unit root

deseason_AF_diff = diff(deseason_AF,differences=1)

adf.test(deseason_AF_diff,alternative="stationary")

deseason_AF_diff_arima_fit=auto.arima(deseason_AF_diff,max.D=0,max.P = 0,max.Q=0)

print(deseason_AF_diff_arima_fit) # We get order (0,0,1) MA

par(mfrow=c(1,1))

plot(dates[2:110],deseason_AF_diff_arima_fit$residuals,type="l", xlab="Month (Jan 2011 - Feb 2020)",
    ylab="Arrivals (# people)",main="Africa Residual")

abline(h=0,col="cyan1",lwd=2)
abline(h=mean(deseason_AF_diff_arima_fit$residuals),col="black",lty="dashed",lwd=1)

par(mfrow=c(1,2))

Acf(deseason_AF_diff_arima_fit$residuals,lag.max=40,main="Africa ACF")

Pacf(deseason_AF_diff_arima_fit$residuals,lag.max=40,main="Africa PACF")
```

Residual Fit

The next step to verify the fit of the ARIMA model is to verify the residual fit of the datasets and to study the residual ACFs and PACFs for the data. Residual fit is what is left of the time series after fitting a model. In a majority of the cases, the residuals are equal to the difference between observations and the corresponding fitted value. Thus, residuals are a good way to judge how good a fit is. Ideally, the residual plot should be evenly distributed around zero. This would mean that our ARIMA fit was able to cover important determinants of the dataset. When it comes to the ACFs and PACFs, majority of the lines should lie within the significance lines. This would mean that it represents a white noise and that the fit represents the data. We also analyzed some of the major highs and lows in the residuals in an attempt to try and understand reasons for significant variations from the mean in some regions.

To compare the previous ARIMA forecast model for Asia's differenced and de-seasonalized time series object of U.S. arrivals, I created a SARIMA model on Asia's original U.S. arrivals dataset by using the *auto.arima()* function available in R statistical programming while in R-Studio to determine the recommended p, d, q, P, D, and Q model parameters. Then, I created an object of the suggested SARIMA model using the *Arima()* function available in R with the recommended parameter values. I then forecasted the next 40 months of U.S. arrivals from Asia based on R's suggested SARIMA model using the *forecast()* function available in R. I plotted this forecast using R's *plot()* function and determined the R-suggested SARIMA model's AIC value using R's *summary()* function. Thereafter, I compared the AIC values from the differenced de-seasonalized ARIMA model and the original data's SARIMA model, by looking for the lower AIC value that would signify a better forecast for U.S. arrivals from Asia.

Below is the R code used to determine the SARIMA model and subsequent AIC value for U.S. arrivals from Asia:

```
library(forecast)
auto.arima(Asia)
arima.asia=Arima(Asia,order=c(0,1,1),seasonal=c(0,1,1))
forecast(arima.asia,h=40)
plot(forecast(arima.asia,h=40),ylab="Arrivals (# of people)",xlab="Year")
summary(arima.asia)
```

Results

In terms of seasonality, not surprisingly, the world regions that had the highest degree of absolute seasonality averaged across 2011 to 2019 were the ones with the highest total number of U.S. arrivals: Western Europe and Asia. Western Europe had the highest degree of absolute seasonality with an average fluctuation of around 180,000 people during a month caused solely by seasonality. Asia had the second highest degree of absolute seasonality with an average fluctuation of around 110,000 people during a month caused solely by seasonality. The lowest degrees of absolute seasonality averaged across 2011 to 2019 were from Africa and Eastern Europe, at fluctuations of around 10,000 people during a month caused solely by seasonality.

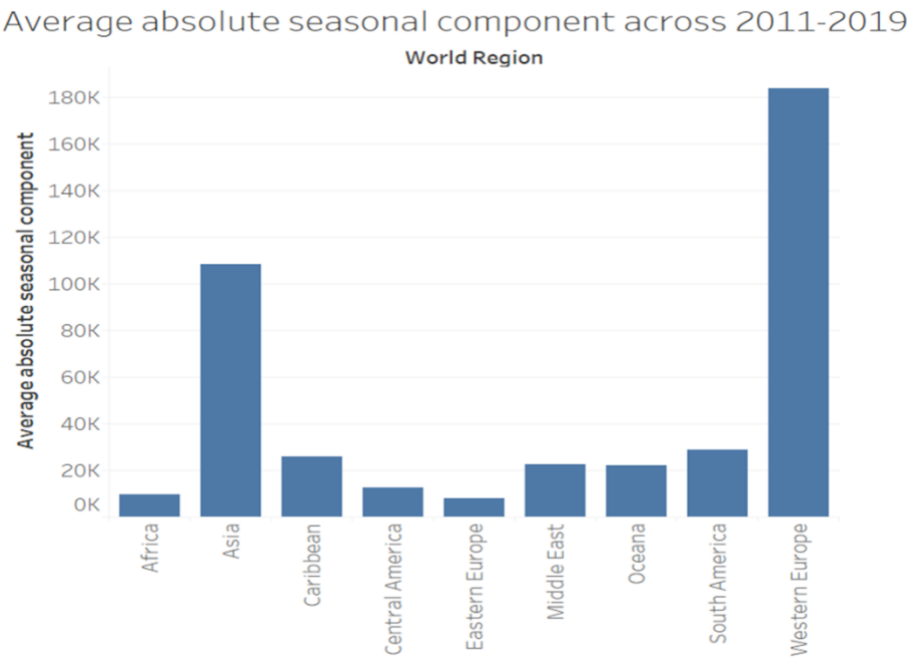

Figure 8: Average absolute seasonal component for each world region averaged across 2011-2019.

World regions with the largest total number of arrivals into the U.S. experienced the greatest fluctuations in monthly U.S. arrivals based on seasonality. This pattern can be clearly seen in a comparison of the original monthly arrivals data of Asia, which has one of the highest overall number of U.S. arrivals, with that of Central America, which has one of the lowest overall number of U.S. arrivals. A larger number of total people arriving into the U.S. from a particular region will make that region's seasonality more pronounced. The larger proportion of U.S. arrivals from Western Europe and Asia is likely due to many factors, some of which include well-established immigration ties to the U.S. and higher socioeconomic status.

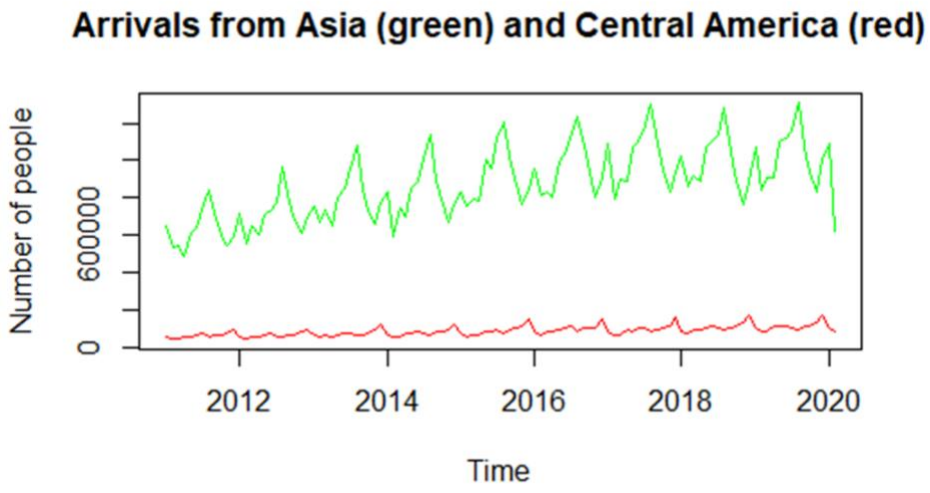

Figure 9: Comparison of arrivals into the U.S. from Asia (green) and Central America (red) from January 2011 to February 2020.

Western Europe's and Asia's strong seasonality continued to prevail when looking at month-specific averages across January 2011 to February 2020. However, even though it does not have a large total number of U.S. arrivals, South America exhibited strong positive seasonality in Decembers. The higher proportion of U.S. arrivals from South America in Decembers is likely due to summer breaks established in this region that allow many people to take time off of work and travel. Overall, the spring month of April tends to bring less people to the U.S. with the strong exception of from Western Europe, the summer month of July tends to bring more people to the U.S., the autumn month of September tends to bring more people to the U.S., and the winter month of December tends to bring less or more people to the U.S depending on the world region.

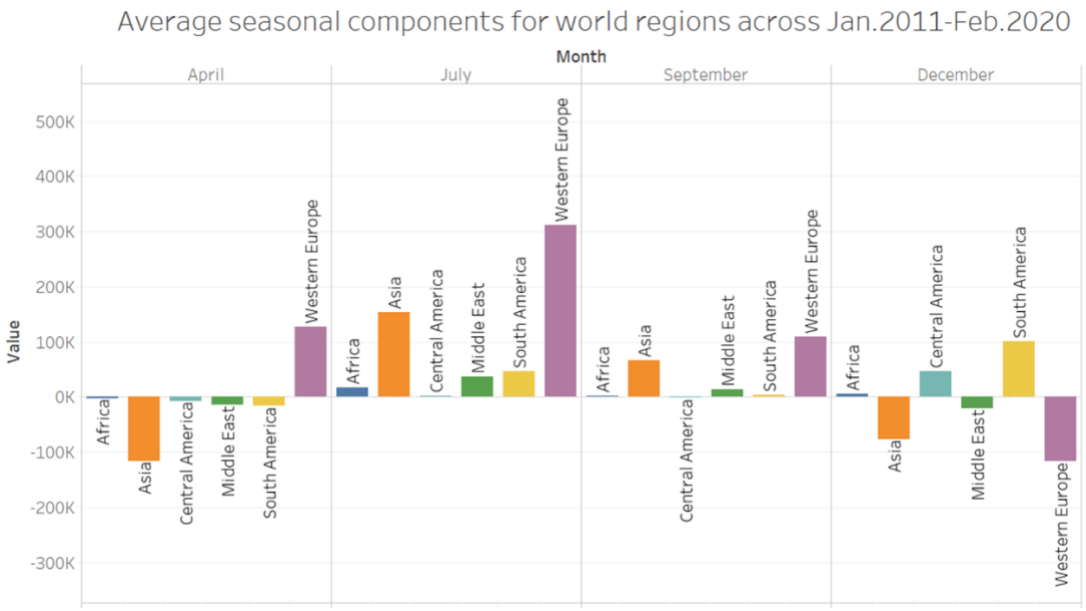

Figure 10: Average seasonal components for world regions across January 2011 to February 2020. Notice the unusually high seasonality from South America in Decembers.

It is also clear in Figure 1 that there is an extreme seasonal component to the time series for every region, with drastic changes occurring annually. There are, in most cases, two peaks in arrivals: summer and the winter around the holidays. Additionally, according to the simple linear regression trendline, there is an increasing number of arrivals to the U.S. from all nine regions. Considering the behavior of the decomposition of the Asia region in Figure 2 above, one clearly visible anomaly in the observed data for the Asian Region decomposition occurs in January and February 2020. During this time period, the arrivals decrease to a level not seen for many years. This is likely due to the COVID-19 pandemic, which struck Southeast Asia before the rest of the world. The early onset of this global disease caused travel disruptions in Asia before they were widespread and seen in other regions. In the deseasoned dataset from Figure 3 and the trend components in Figure 4, the significant variation previously seen is reduced, which highlights certain anomalies which were previously difficult to detect with the extreme seasonality present in the time series. For example, in Western Europe and Oceana, there are significant increases in the number of arrivals during 2014 and 2015, followed by a sharp decrease back to the anticipated trend in 2016.

The comparison between arrivals from Western Europe and the value of the USD (Figure 5) shows no correlation that would explain increased arrivals in the U.S. Although the value does increase simultaneously, this would actually have an inverse effect on the number of arrivals as it would cost relatively more to visit. Therefore, the value of the USD does not function as a significant explanatory variable for the number of arrivals. The other potentially impactful factor that would lead to the decrease in travel to the U.S. after 2016 is the 2016 Presidential Election, where President Donald Trump was elected. Throughout his campaign for president, he repeatedly attacked other countries and threatened to “close the border.” Despite not having enacted any policies regarding this issue prior to the visible decline in arrivals, the presence of

the rhetoric likely impacted the attitudes of individuals in other countries towards the U.S., reducing the individuals arriving in the country.

On trying to determine the fit of the de-seasonalized differenced ARIMA model, we noticed a low AIC value for all the datasets, which indicated a good fit. Afterwards, we got the perfect fit for each model as shown in table 1. We can see that in contrast to the earlier assumption that most of the datasets follow an MA trend.

Table 1: comparison of residual plots for all nine regions.

Region	Order	Fit	AIC
Western Europe	(0,1,2)	MA	2702.24
Asia	(0,0,2)	MA	2645.97
South America	(0,0,1)	MA	2491.74
Caribbean	(0,1,1)	MA	2335.8
Oceana	(2,0,0)	AR	2210.44
Central America	(0,1,1)	MA	2237.07
Middle East	(0,0,1)	MA	2366.48
Eastern Europe	(1,0,1)	ARMA	2127.77
Africa	(0,0,1)	MA	2085.29

We see that majority of the residual plots average out at zero, being distributed evenly around 0. We also see that some of the plots show a significant variation from the regular trend at some point of time. Three of the regions that show this variance are Africa and Caribbean in 2012 and Asia in 2020. Their residuals are plotted in figure 11. We can see that South Africa and Caribbean show a major drop in travel to the United States during early 2012. On investigation about the cause of this issue, we found two possible reasons for this. In Africa, 2012 was the year of the start of Central African Republic Civil War. While considering the Caribbean region, one possible reason for the drop in visitors to US can be hurricane Sandy, which hit United States in 2012 but also affected parts of Bermuda and the Caribbean. Lastly, looking at Asia, there is a significant drop in visitors to the United States in early 2020. A clear reasoning for this trend is, again, coronavirus hitting China. Majority of visitors to United States from Asia are from China. Since China was the first country to be hit by coronavirus and the first to impose lockdown and restriction to travel, it is evident that this led to a drop in visitors in early 2020. It will also be interesting to see this effect pan out as coronavirus spreads across the globe.

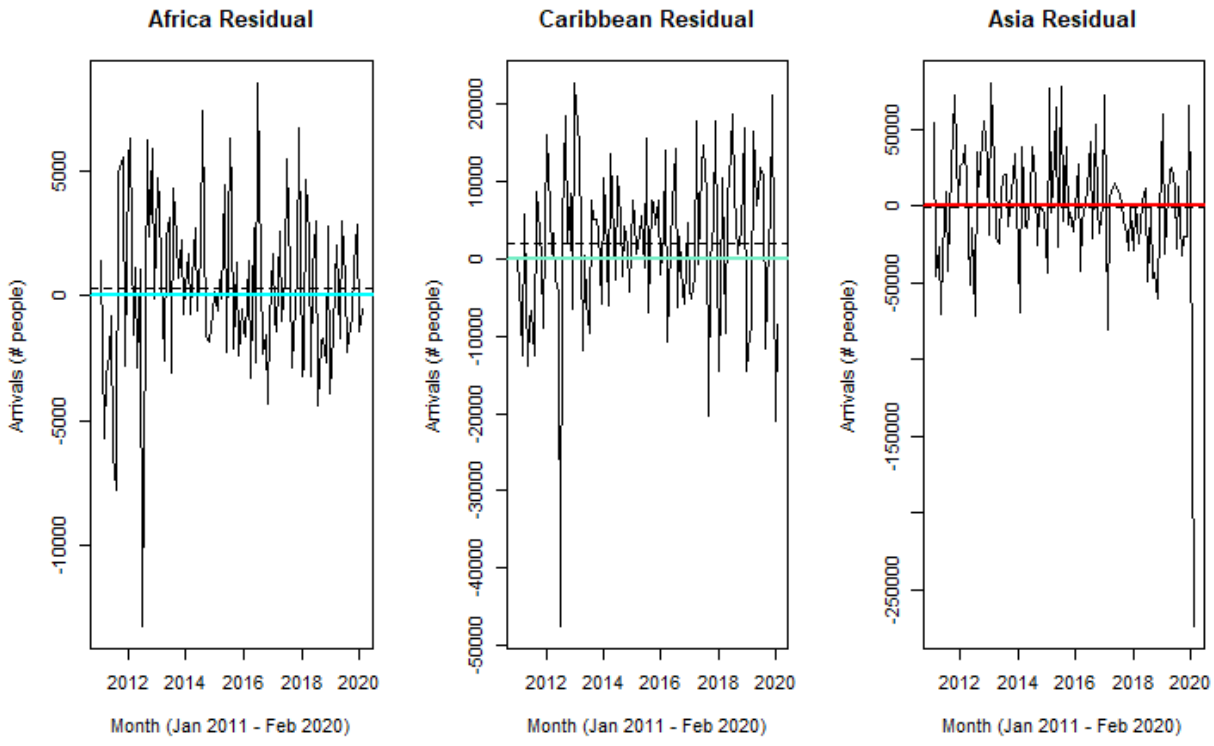

Figure 11: Residual comparison for Africa, Caribbean and Asia.

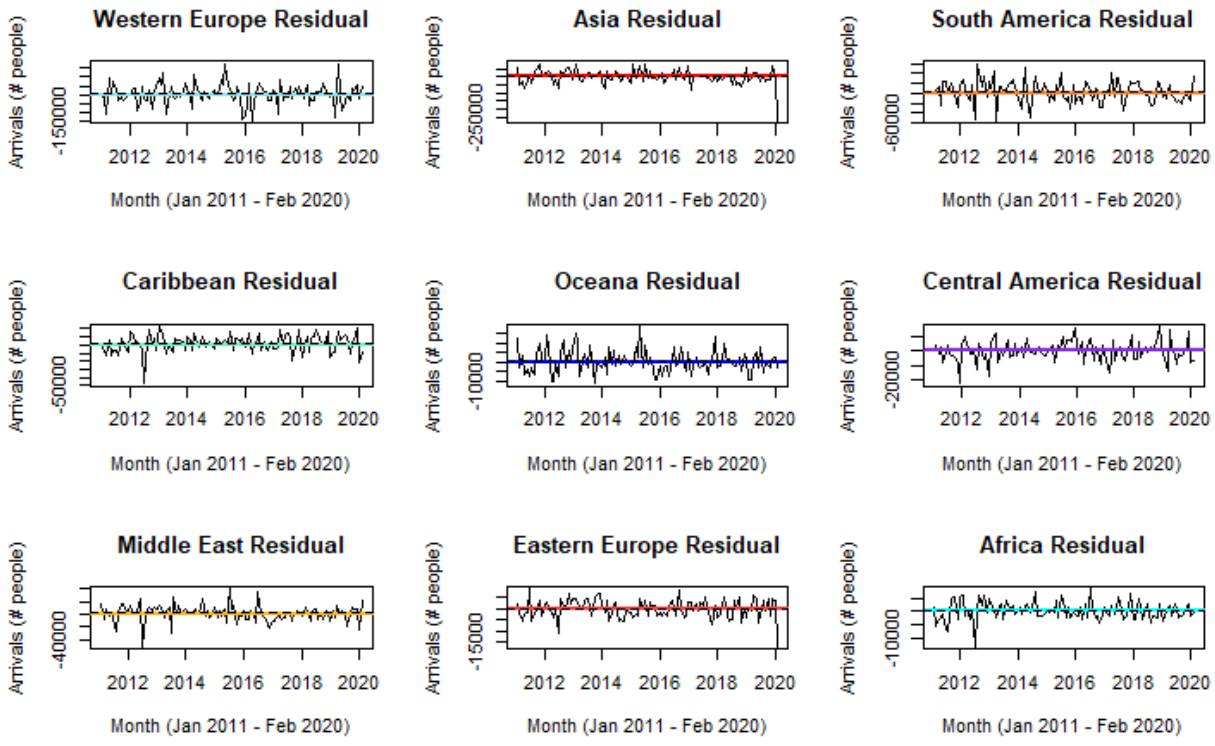

Figure 12: Residual plots for all regions

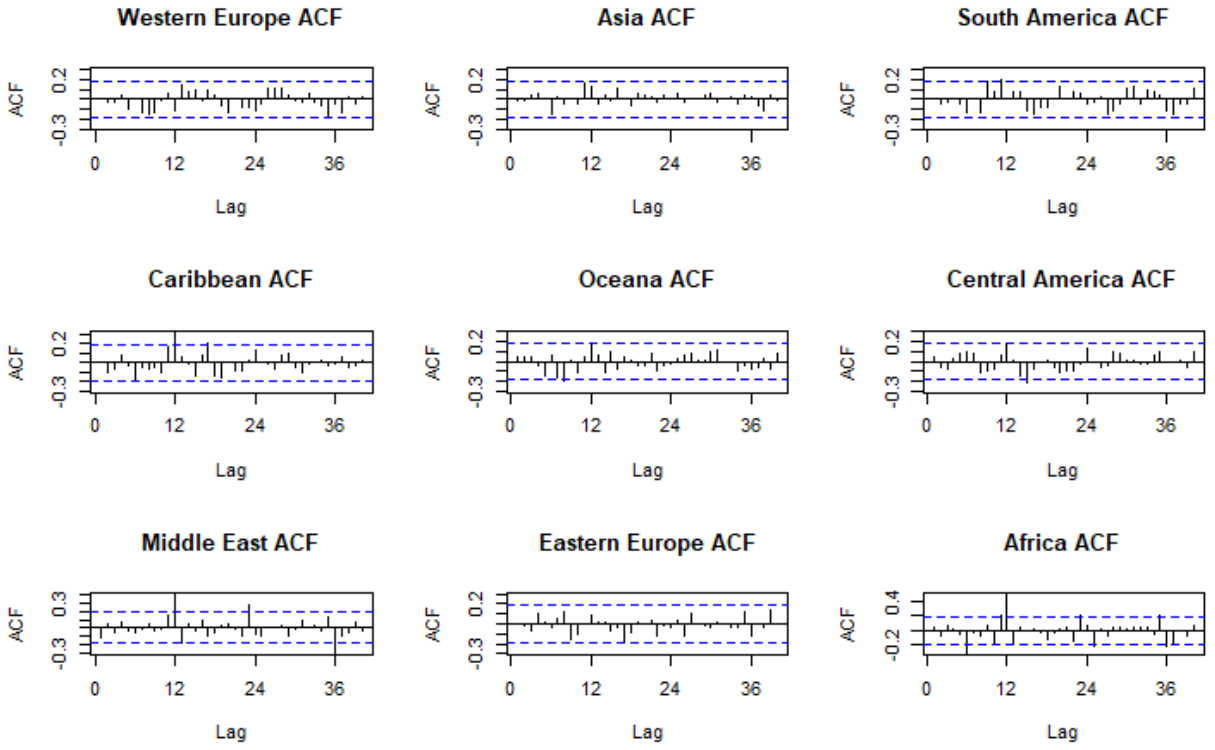

Figure 13: Residual ACFs for all regions

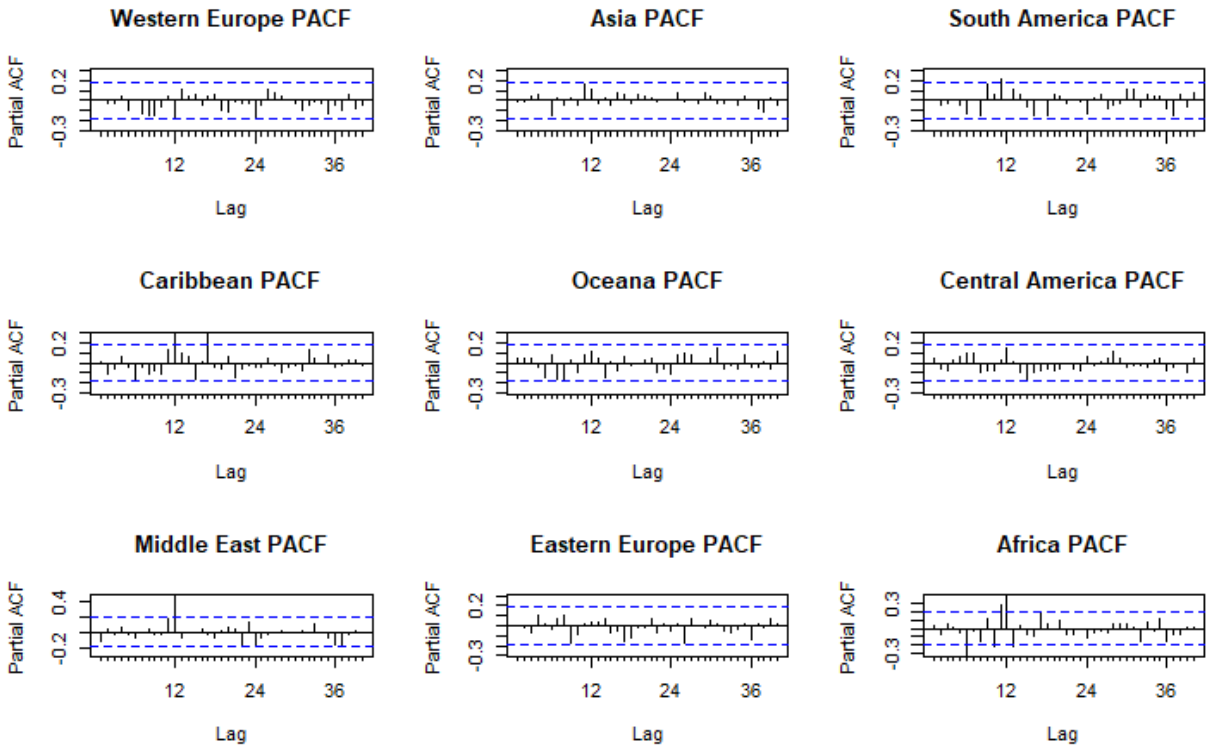

Figure 14: Residual PACFs for all regions

Figure 13 shows the residual ACFs for all regions and figure 14 shows the residual PACFs for all regions. Looking at the ACFs and PACFs we can infer that a majority of the lines lie within the significance line except for the select few discrepancies that we talked about earlier. This helps us conclude that the residuals represent white noise post fitting. This further strengthens our assumption that our ARIMA fit is accurate for all regions and is appropriate for forecasting and making future predictions. The presence of white noise in the ACF and PACF, along with our data being deseasonal also helps us conclude that seasonality is a major determinant of the dataset and thus will play an important role in forecasting.

To compare the previous de-seasonalized differenced ARIMA model, R suggested the following SARIMA model parameters for the original U.S. arrivals from Asia data: p=0, d=1, q=1, P=0, D=1, Q=1. The subsequent AIC value for this SARIMA model is 2,372 and is lower than the 2,646 AIC value of the previous ARIMA model from the differenced de-seasonalized U.S. arrivals from Asia data. Therefore, the SARIMA model of the original data is the better forecast for U.S. arrivals from Asia.

Conclusion

Based on I-94 forms, arrivals into the United States during the past decade have steadily increased overall and fluctuated strongly with the seasons. Because of the data's strong seasonality and relatively steady increasing trend, both the de-seasonalized ARIMA as well as the SARIMA models do a very good job at forecasting. However, major events such as COVID-19, Ebola outbreak of 2012, U.S. dollar valuation changes, and many others can be sudden, have drastically altered arrival patterns in the past decade, and are very hard to incorporate into forecast models. Therefore, users of these models can project future monthly arrivals into the U.S. from each world region for the next couple years with relatively high degree of confidence, but should also be prepared for sudden events that can significantly alter U.S. arrivals trends from certain regions or even worldwide.

In addition, it is important to note some significant limitations of working with the I-94 dataset. Firstly, the models we proposed in this project are not representative of immigration patterns since arrivals does not equate to immigration. Whereas immigration refers to long-term residence in a country, arrival can arise from many other motives such as tourism, visiting relatives during holidays, and business trips. For example, even though the project's models indicate a continuously increasing trend of arrivals into the U.S. during the past decade, news sources indicate that immigration into the U.S. has actually stalled during the past decade (*Times Record News*, 2020). Secondly, the analyzed dataset does not include the vast majority of U.S. arrivals from Mexico nor Canada, since most citizens from those two countries are not required to fill out the form. A large majority of legal immigration to the U.S. comes from Mexico, as can be seen in the below map from 2012, and hence the dataset is severely lacking in representation of these large arrival numbers.

Figure 15: U.S. map of the most common country of origin of legal immigrants for each state in 2012.

References

"Form I-94, Arrival/Departure Record, Information for Completing USCIS Forms." *USCIS*, 3 July 2018, www.uscis.gov/I-94information.

"Frequently Asked Questions." *I-94 Official Website*, US Customs and Border Protection, i94.cbp.dhs.gov/I94/#/faq.

"I-94 Home." *I-94 Official Website*, US Customs and Border Protection, i94.cbp.dhs.gov/I94/#/home.

"International Visitation and Spending in the United States." *National Travel and Tourism Office (NTTO)*, US Department of Commerce, travel.trade.gov/outreachpages/inbound.general_information.inbound_overview.asp.

Kowalick, Claire. "Net Immigration to U.S. Lowest in a Decade, China Surpasses Mexico for Incoming Migrants." *Times Record News*, Jan.3 2020, https://www.timesrecordnews.com/story/news/local/2020/01/03/net-immigration-us-lowest-decade-china-tops-migrants/2804697001/ .

"U.S. Dollar Index - 43 Year Historical Chart." *MacroTrends*, www.macrotrends.net/1329/us-dollar-index-historical-chart.