Eindopdracht - Predictive Modeling: Forecasting

Joshua de Freitas

2022-17-11

Introduction

Remittances to Mexico reached a record high in July, with Mexican families receiving $5.3 billion from abroad, an increase of 16.5% compared to the previous year (Banxico, 2022). For many developing economies, such as Mexico and the CADPR region (Central America, Panama, and the Dominican Republic), remittances are a vital source of funds, often surpassing official aid or foreign direct investment. Remittances can be defined as income received by households from foreign economies, primarily arising from the temporary or permanent movement of workers to those economies. This income can include cash, as well as non-cash items sent or given through formal channels, such as electronic cash transfers, or through informal channels, such as money or goods taken across borders (IMF, 2009).

For this assignment, a univariate time series dataset from the Mexican central bank *Banco de Mexico* will be used to build forecasting models and make predictions about the amounts of future remittance that are received in Mexico. The dataset includes 334 *monthly* observations from January 1995 to August 2022, with the target variable **remittances** measured in millions of USD. An initial visualization of the data is shown in Figure 1 below.

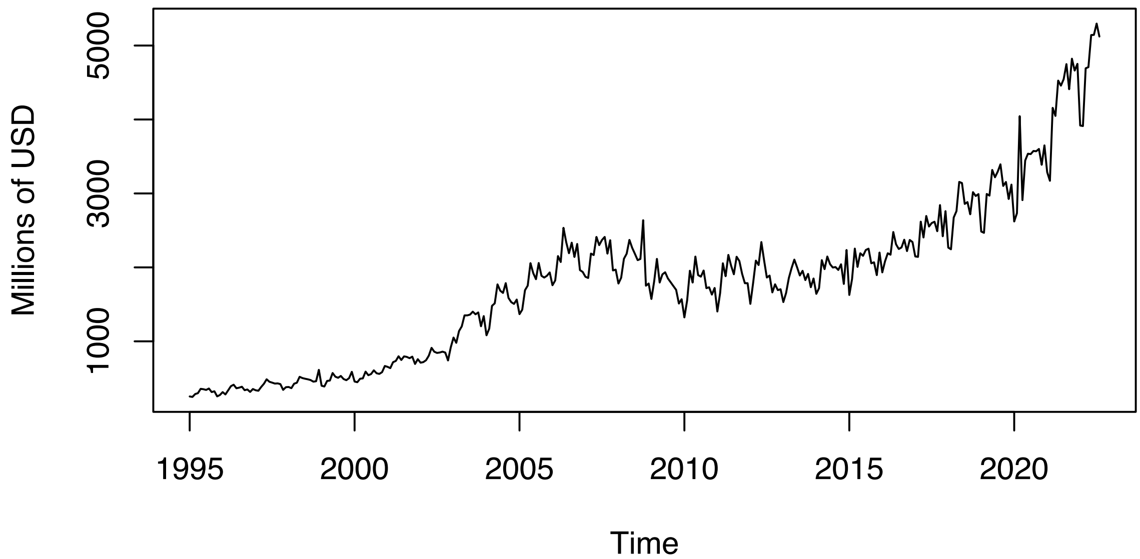

Figure 1: Remittances received in Mexico by month (Jan. 1995 to Aug. 2022)

Methodology

To forecast future remittances to Mexico, the following steps will be taken:
1. The time series will be split into a training set and a test set. The training set will be used to estimate the parameters of the forecasting model, and the test set will be used to evaluate its predictions.
2. Plots of the autocorrelation function (ACF) and partial autocorrelation function (PACF) will be used to analyze the underlying trends and seasonal patterns of the series. These plots will also be used to identify the appropriate lag structure or order of the model terms.
3. Trends and seasonal patterns will be removed to make the time series stationary, after which the appropriate model terms will be identified using the ACF and PACF plots.
4. A first model will be estimated based on the training set and compared against alternative models. The best model will be chosen according to the relative quality measures and model accuracy measures.
5. The predictive power and accuracy of the final model will be evaluated on the test set, after which the final model will be re-estimated using the original time series to make future predictions for a period of 4 years (48 months).

1. Splitting the time series: Training and test set

As previously mentioned, the first step in constructing a model is to split the time series into a training set and a test set. The training set will be used to estimate the parameters of the forecasting model, and the test set will serve to evaluate its predictions. The following code performs this split, resulting in a training set containing **284** observations and a test set containing **48** observations. The test set therefore includes the final 48 months of the time series, allowing for the evaluation of the model's ability to forecast future values.

```
test_size <- 48
data_size <- length(Y)
training_size <- data_size - test_size
training <- head(Y, n = training_size)
test <- tail(Y, n = test_size)
```

2. Analysis of trends and seasonality

The previous graph in figure (1) showed a clear upward and exponentially growing trend in the time series as well as repeating seasonal patterns. While its momentum stagnated between 2007 and 2010, perhaps due to the impact of the global and US financial crisis, the exponential growth of the time series appears to be almost continuous. Furthermore, the variation of the time series also appears to be growing over time. These observations indicate that the time series does not appear to be stationary in both its mean and variation. This can also be analyzed more quantitatively using the *Augmented Dickey-Fuller Test*, as shown below.

```
    Augmented Dickey-Fuller Test

data:  training
Dickey-Fuller = -1.1258, Lag order = 12, p-value = 0.917
alternative hypothesis: stationary
```

The results of the ADF-test show that the p-value (0.917) is larger than the 5% significance level, suggesting that the time series variable is indeed non-stationary.

Since, many econometric forecasting models assume that the data on which they are applied is stationary (i.e. having a constant mean and variance), ACF and PACF plots are used to more closely analyze the underlying trends and seasonal patterns of the series.

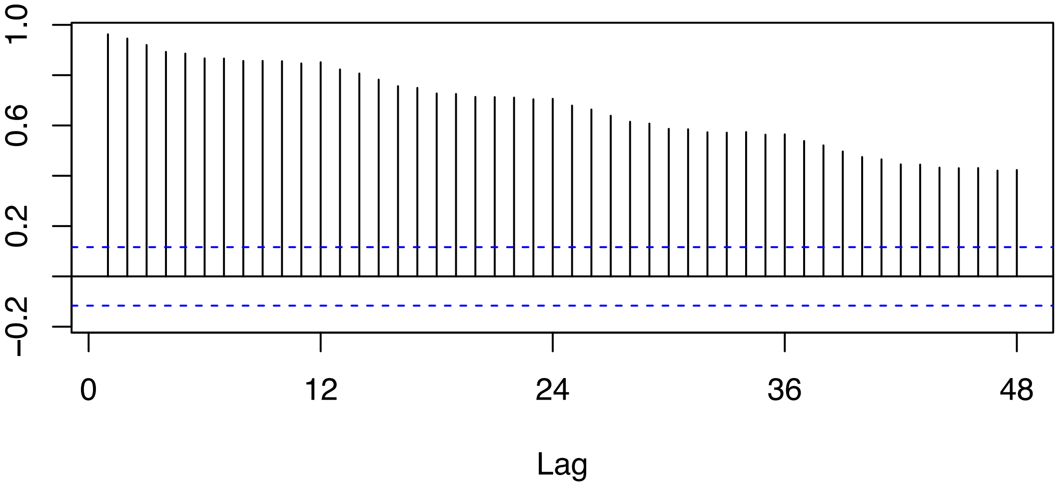

Figure 2: Autocorrelation function (ACF)

The autocorrelation function or ACF in figure (2) shows the degree of correlation between the time series and its lagged values and it helps to identify any significant autocorrelations that may be present in the data. In this case, the ACF decays slowly, suggesting that the series exhibits a long-term trend. The series also appears to spike each time after 12 lags, which likely indicates that there is a recurring annual seasonal pattern in the data. It will be important to account for these patterns in the next step of the analysis.

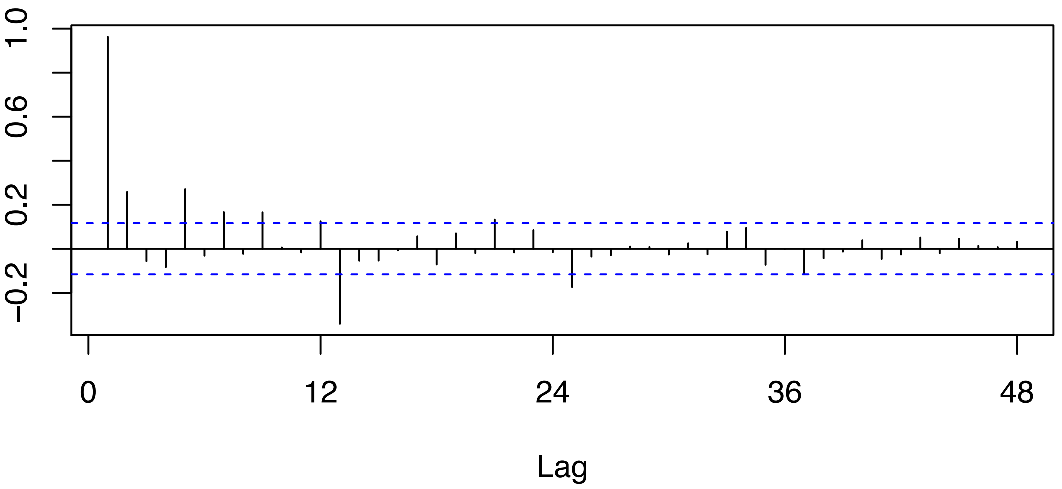

Figure 3: Partial autocorrelation function (PACF)

The partial autocorrelation function (PACF) is a measure of the correlation between a variable and its lagged values, taking into account the correlations of other lagged values. It can be used to identify the unique correlation between a variable and each individual lag, allowing for the detection of patterns and trends in the data that can be accounted for by incorporating these lags into a model. Figure (3) above shows significant correlations at the 1st, 2nd, and 5th lags, which may be used in modeling the data.

3. Transformations of the time series

As described in the previous section, the time series is clearly non-stationary as it exhibits a strong upward trend as well as recurring (annual) seasonality, as was shown in figure (1). In order to make the time series more stationary, a log-transformation of the series is first applied to stabilize its variation. This is done in order to prevent biased or incorrect estimates.

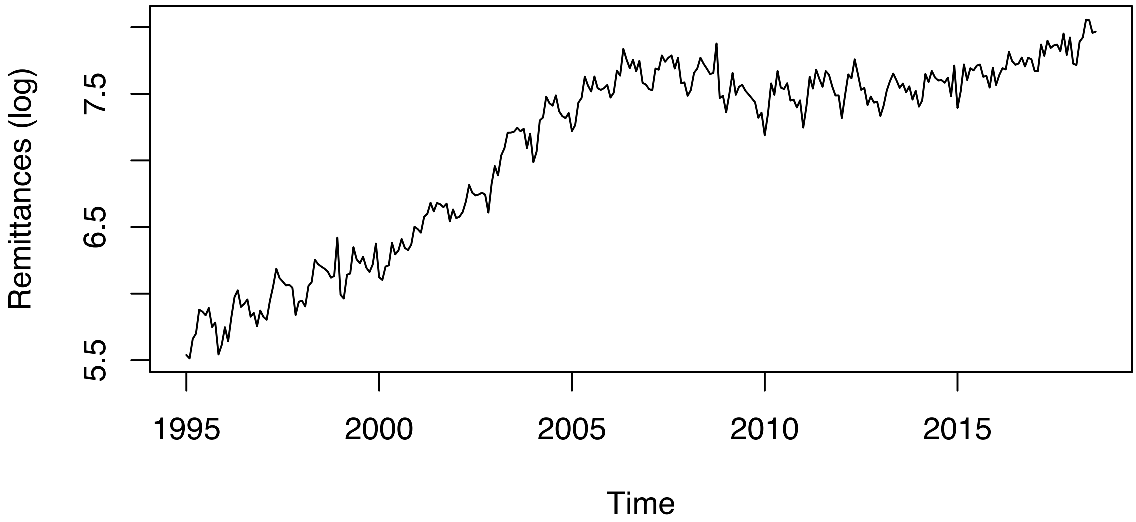

Figure 4: Log remittances received in Mexico by month

Taking the log of the time series now reveals a more linear and stable variation, which is consistent with our desired expectation. The next step involves taking a first difference, meaning that the changes in monthly values are used to construct a model instead of the actual individual observations. As a result, the long-term effects of the identified trend are removed, which in turn helps to reveal the underlying pattern of the series as well as highlight any short term fluctuations.

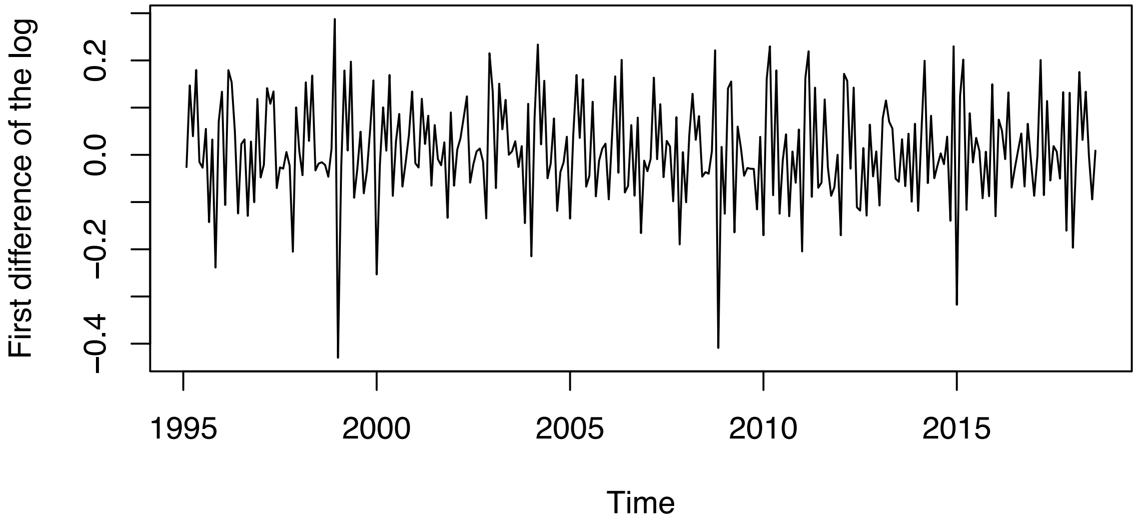

Figure 5: First difference of the log remittances received in Mexico by month

In addition to removing the trend in the time series, it is also necessary to account for the observed seasonality patterns. One way to achieve this is by taking a seasonal difference on top the initial first difference. As a result, the influence of recurring seasonality is effectively accounted for. This is shown in figure (5) below.

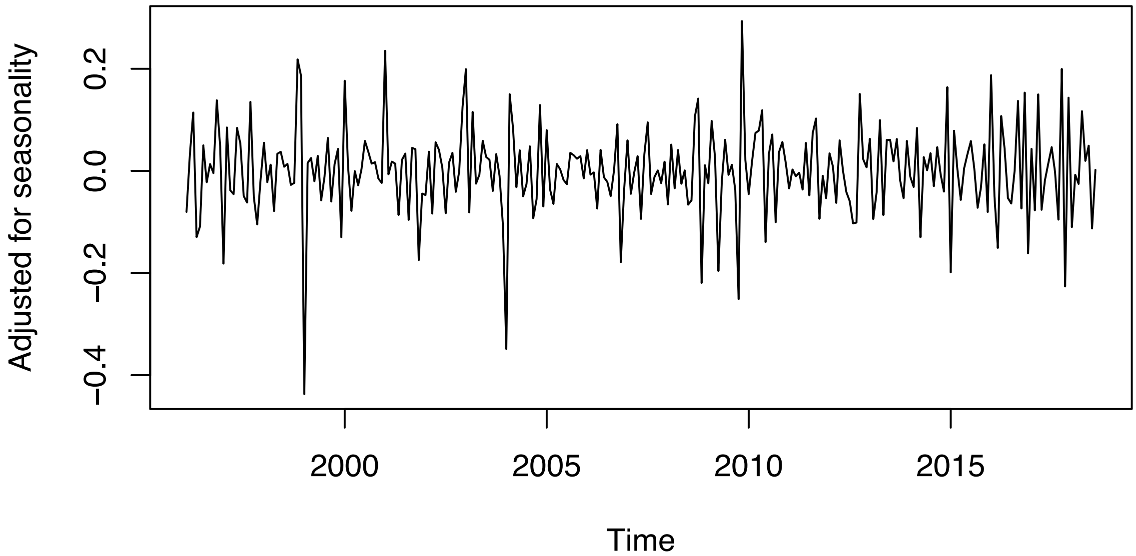

Figure 6: First difference of the log remittances received in Mexico by month

After performing these transformations, the time series now appears to be stationary. This is also confirmed by the ADF-test, where the P-value (0.01) is now much lower than the 5% significance level, as shown below.

```
	Augmented Dickey-Fuller Test

data:  training.final
Dickey-Fuller = -6.949, Lag order = 12, p-value = 0.01
alternative hypothesis: stationary
```

4. Model identification and comparison

Now that the time series has been made stationary, an important assumption for using ARIMA-models, a first model can be constructed to fit the training data. To do this, ACF and PACF plots are analyzed to identify the appropriate order of the model terms (p, d, q), where **p** refers to the autoregressive process(es), **d** to the number of differences that have been applied, and **q** to the moving average process(es). However, these terms refer to the *non-seasonal* factors of the ARIMA model. To extend this model, seasonal components, **P**, **D** and **Q** are included, where P and Q incorporate the seasonal factors and D refers to the number of seasonal differences that have been applied. Given that a first difference as well as a seasonal difference have been applied on the series, both d and D are set to 1.

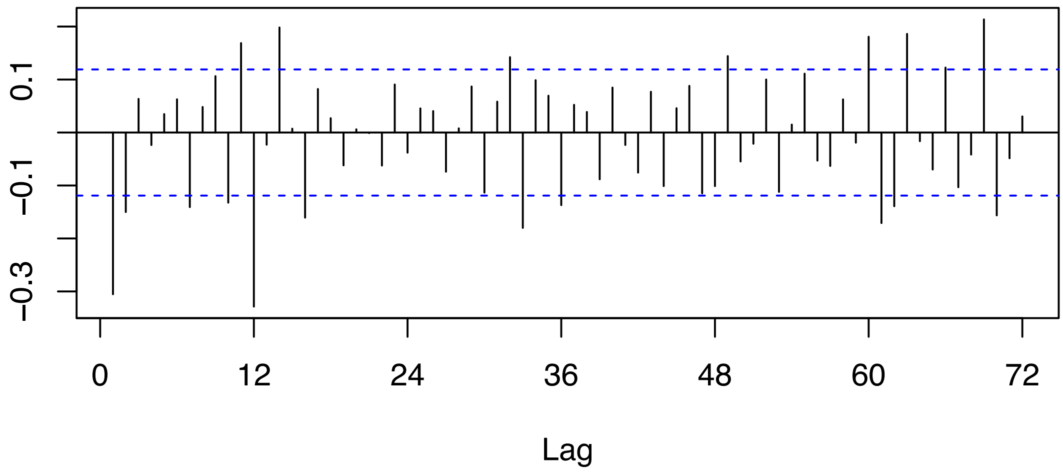

Figure 7: Autocorrelation function (ACF)

The ACF plot shows significant autocorrelations at the 1st and 2nd lag, suggesting either a non-seasonal MA(1) or MA(2) process. The significant spike at the 12th lag may suggests a seasonal MA(1) component because the time series contains monthly intervals.

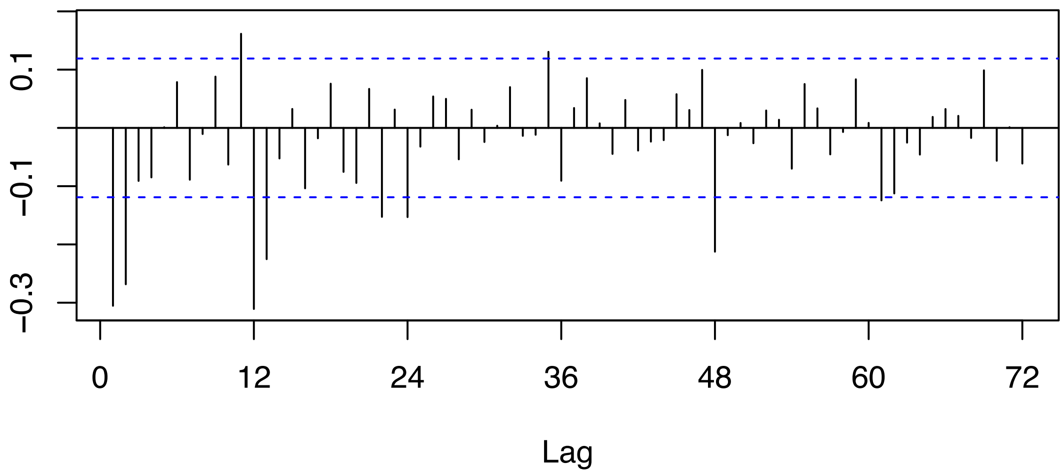

Figure 8: Partial autocorrelation function (PACF)

Similarly, the PACF plot also shows significant autocorrelations at the 1st and 2nd lag as well as at the 12th lag, indicating a non-seasonal AR(1) or AR(2) process and a seasonal AR(1) process. Based on these observations, the resulting ARIMA-model can thus be identified as **ARIMA(2, 1, 2)(1, 1, 1)12**.

The summary statistics for this seasonal ARIMA-model below show the estimated coefficients and standard errors of the model parameters along with the measures of the relative quality of model (AIC, AICc and BIC). The lower these scores, for instance in the case of the AIC-metric, the better the model is able to capture or fit the training data.

```
Series: training
ARIMA(2,1,2)(0,1,1)[12]

Coefficients:
          ar1     ar2     ma1      ma2     sma1
      -0.9013  0.0187  0.4141  -0.5664  -0.5887
s.e.   0.1090  0.1084  0.0893   0.0889   0.0661

sigma^2 = 12323:  log likelihood = -1661.87
AIC=3335.75   AICc=3336.06   BIC=3357.36
```

In this case, the AIC is equal to **3336.06**. To compare this result against other model specifications, three alternative models are estimated below:

```
# Comparison of Alternative models based on relative quality measure (AIC)

# ARIMA(0,1,1)(1,1,1)12
m2 <- Arima(training, order = c(0, 1, 1), seasonal = c(1 ,1, 1))
c('AIC:', round(AIC(m2), 2))
```

```
[1] "AIC:"    "3335.17"
```

```
# ARIMA(0,1,2)(0,1,1)12
m3 <- Arima(training, order = c(0, 1, 2), seasonal = c(0, 1, 1))
c('AIC:', round(AIC(m3), 2))
```

```
[1] "AIC:"    "3337.75"
```

```
# ARIMA(1,1,1)(0,1,1)12
m4 <- Arima(training, order = c(1, 1, 2), seasonal = c(0, 1, 1))
c('AIC:', round(AIC(m4), 2))
```

```
[1] "AIC:"    "3333.78"
```

Based on a comparison of the AIC-metric, the last model **ARIMA(1,1,2)(0,1,1)12** provides the best fit to the training data. In addition to comparing the relative quality of the models using the AIC, a comparison is made based on the model accuracy measures (ME, RMSE, MAE, MPE, MAPE, MASE, ACF1).

```
[1] "Model accuracy measures ARIMA(2,1,2)(0,1,1)12"

                   ME    RMSE      MAE       MPE   MAPE      MASE         ACF1
Training set 27.19311 203.186 126.6006 0.4604208 3.2681 0.2175451 -0.007695591

[1] "Model accuracy measures ARIMA(0,1,1)(1,1,1)12"

                  ME     RMSE      MAE      MPE     MAPE      MASE
Training set 23.3892 207.3133 130.2609 0.375848 3.378784 0.2238346
                      ACF1
Training set -0.0004573412
```

```
[1] "Model accuracy measures ARIMA(0,1,2)(0,1,1)12"

                   ME     RMSE     MAE       MPE     MAPE      MASE        ACF1
Training set 27.73204 205.1587 128.406 0.4686051 3.311573 0.2206474 -0.01619573

[1] "Model accuracy measures ARIMA(1,1,1)(0,1,1)12"

                   ME     RMSE      MAE       MPE     MAPE      MASE
Training set 26.92838 203.3478 126.7512 0.4544382 3.272586 0.2178038
                     ACF1
Training set -0.001917979
```

A comparison of the accuracy measures also reveals that **ARIMA(1,1,2)(0,1,1)12** is the preferred model, since it provides the most accurate predictions on the test set out of all four models according to these measures.

The resulting summary statistics of this model, i.e. the estimated coefficients and standard errors along with the relative quality measures and model accuracy measures are again shown below in more detail:

```
Series: training
ARIMA(1,1,2)(0,1,1)[12]

Coefficients:
          ar1     ma1      ma2     sma1
      -0.9192  0.4264  -0.5538  -0.5874
s.e.   0.0323  0.0557   0.0531   0.0658

sigma^2 = 12279:  log likelihood = -1661.89
AIC=3333.78   AICc=3334   BIC=3351.79

Training set error measures:
                   ME     RMSE      MAE        MPE     MAPE      MASE
Training set 3.524386 107.4411 74.95548 0.06813488 4.999407 0.4572488
                    ACF1
Training set 0.002429527
```

The following graph in figure (9) plots the fitted values of the model (blue line) against the original observations (black line).

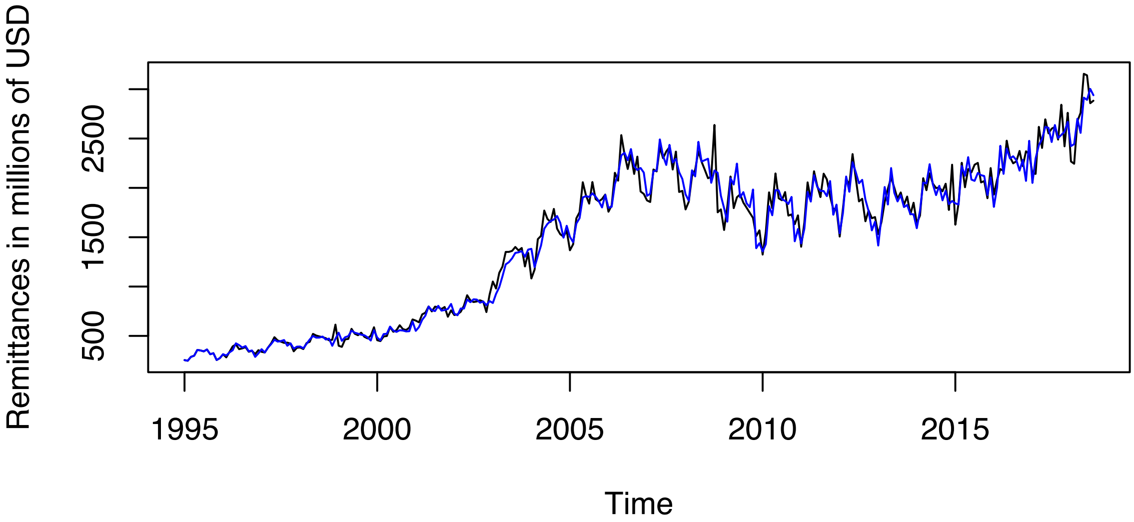

Figure 9: Fitted values vs. Original observations

While the model does not provide a perfect fit to the data, which is also undesirable due to the problem of *overfitting*, it does decently capture the overall trend of the time series as well as the short-term fluctuations within each year. The next graph in figure (10) plots the predicted values and confidence intervals of the remaining 48 months of the series (in blue) along with the actual values of the test set (in orange).

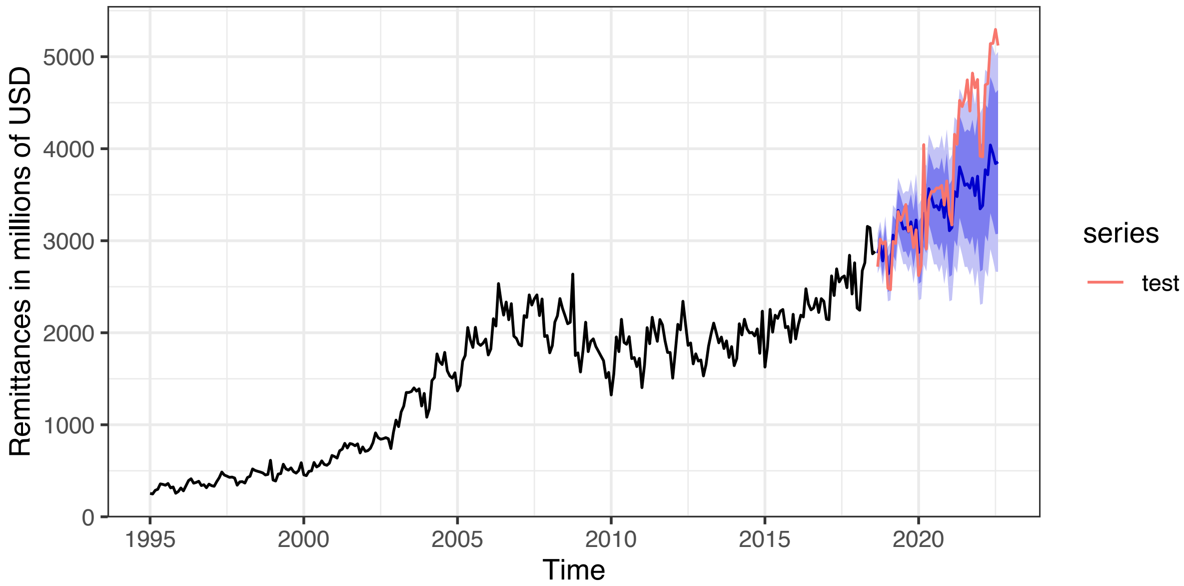

Figure 10: Predicted values based on the training set (h=48)

It is clear that the model is well able to make predictions on the first few months of the test set as the blue and orange lines are very close to on another and the original observations fall within the range of the confidence intervals of the forecasts. However there is a sudden strong increase in the amount of remittances in the year 2021, after which the model predictions and the observations in the test set begin to grow more apart. Nevertheless, the model can be regarded as a useful model to make accurate predictions. The model can now be used to make forecasts of future values after it has been re-estimated on the entire dataset, which will be done in the next step.

5. Forecasting future values

The model will now be re-estimated on the entire dataset to make predictions on future values, namely on the upcoming 48 months. The resulting summary statistics of the re-estimation of the **ARIMA(1,1,2)(0,1,1)12** model are shown below.

```
Series: Y
ARIMA(1,1,2)(0,1,1)[12]

Coefficients:
          ar1      ma1      ma2     sma1
      -0.3815  -0.1235  -0.2730  -0.6046
s.e.   0.5581   0.5458   0.2856   0.0513

sigma^2 = 18760:  log likelihood = -2022.93
AIC=4055.87   AICc=4056.06   BIC=4074.69

Training set error measures:
                   ME     RMSE      MAE       MPE     MAPE      MASE
Training set 8.973992 133.4136 88.80925 0.1576708 4.966305 0.4126508
                     ACF1
Training set -0.005648122
```

The next graph plots the predictions on future values (i.e. values after August 2022) based on the entire dataset.

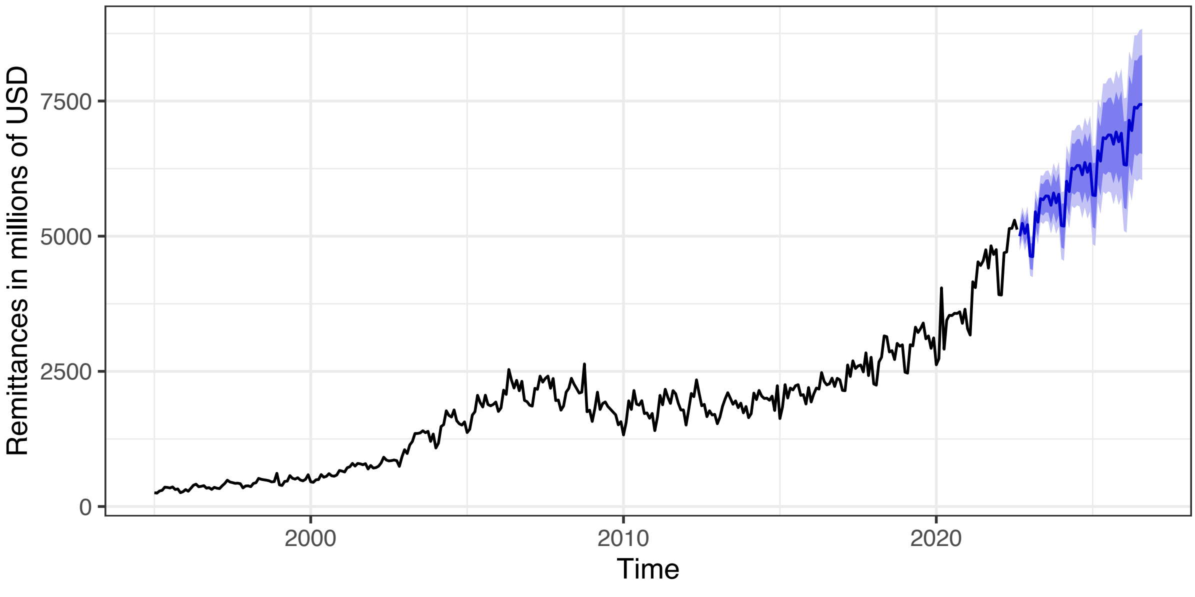

Figure 11: Predicted values based on the complete set (h=48)

The model predictions on the future 48 months are depicted by the blue line along with their 95% confidence intervals (light blue range) and 80% confidence intervals (dark blue range). The actual point forecasts for each future value along with their 95% and 80% confidence intervals are shown below.

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
Sep 2022	4998.581	4823.052	5174.111	4730.132	5267.031
Oct 2022	5238.213	5042.354	5434.073	4938.673	5537.754
Nov 2022	5051.534	4842.590	5260.479	4731.981	5371.087
Dec 2022	5212.751	4989.665	5435.837	4871.571	5553.931
Jan 2023	4630.479	4394.767	4866.191	4269.989	4990.969
Feb 2023	4621.368	4373.431	4869.305	4242.181	5000.554
Mar 2023	5452.262	5192.763	5711.760	5055.393	5849.130
Apr 2023	5259.646	4989.047	5530.244	4845.801	5673.491
May 2023	5695.038	5413.789	5976.287	5264.904	6125.172
Jun 2023	5672.735	5381.219	5964.250	5226.900	6118.569
Jul 2023	5744.522	5443.092	6045.953	5283.524	6205.521
Aug 2023	5742.125	5431.094	6053.155	5266.445	6217.804
Sep 2023	5571.156	5227.529	5914.782	5045.624	6096.687
Oct 2023	5799.364	5438.244	6160.483	5247.079	6351.648
Nov 2023	5617.042	5240.840	5993.244	5041.691	6192.393
Dec 2023	5776.597	5385.314	6167.879	5178.182	6375.011
Jan 2024	5194.959	4789.371	5600.547	4574.665	5815.252
Feb 2024	5185.606	4766.120	5605.091	4544.058	5827.153
Mar 2024	6016.592	5583.685	6449.499	5354.518	6678.666
Apr 2024	5823.940	5378.004	6269.877	5141.940	6505.941
May 2024	6259.346	5800.755	6717.937	5557.991	6960.701
Jun 2024	6237.038	5766.130	6707.946	5516.847	6957.229
Jul 2024	6308.828	5825.918	6791.737	5570.281	7047.374
Aug 2024	6306.429	5811.808	6801.050	5549.972	7062.886
Sep 2024	6135.460	5610.327	6660.594	5332.338	6938.583
Oct 2024	6363.668	5819.844	6907.493	5531.960	7195.376
Nov 2024	6181.347	5620.830	6741.864	5324.110	7038.583
Dec 2024	6340.901	5763.671	6918.132	5458.103	7223.699
Jan 2025	5759.264	5165.978	6352.550	4851.911	6666.616
Feb 2025	5749.910	5140.922	6358.898	4818.544	6681.277
Mar 2025	6580.897	5956.628	7205.166	5626.160	7535.634
Apr 2025	6388.245	5749.050	7027.440	5410.681	7365.809
May 2025	6823.651	6169.875	7477.427	5823.787	7823.515
Jun 2025	6801.343	6133.302	7469.383	5779.663	7823.022
Jul 2025	6873.132	6191.126	7555.138	5830.094	7916.170
Aug 2025	6870.734	6175.042	7566.425	5806.765	7934.702
Sep 2025	6699.765	5974.012	7425.518	5589.821	7809.708
Oct 2025	6927.973	6182.193	7673.752	5787.401	8068.544
Nov 2025	6745.651	5981.601	7509.701	5577.138	7914.165
Dec 2025	6905.206	6122.857	7687.555	5708.707	8101.705
Jan 2026	6323.568	5523.510	7123.627	5099.984	7547.152
Feb 2026	6314.215	5496.766	7131.663	5064.035	7564.394
Mar 2026	7145.201	6310.749	7979.653	5869.017	8421.386
Apr 2026	6952.550	6101.425	7803.675	5650.866	8254.233
May 2026	7387.955	6520.481	8255.430	6061.268	8714.643
Jun 2026	7365.647	6482.125	8249.170	6014.416	8716.878
Jul 2026	7437.437	6538.153	8336.720	6062.101	8812.772
Aug 2026	7435.038	6520.265	8349.812	6036.013	8834.064

References

Banxico. (2022). *Revenues by workers' remittances - (CE81)*. Banco de Mexico. https://www.banxico.org.mx/SieInternet/consultarDirectorioInternetAction.do?accion=consultarCuadro&idCuadro=CE81&locale=en

IMF. (2009). *Balance of payments and international investment position manual*. INTERNATIONAL MONETARY FUND. https://www.imf.org/external/pubs/ft/bop/2007/pdf/bpm6.pdf