Modelling tourism occupancy in Puerto Rico: a neural network approach

First Draft

Marta Álvarez, Ph.D.
Professor
Institute of Statistics and Computerized Information Systems
School of Business Administration
University of Puerto Rico
Río Piedras Campus
PO Box 23332
San Juan, PR 00931
Tel. 787.764.0000, Ext. 87065, 87066 y 87099
marta.alvarez1@upr.edu

Abstract

In this paper an artificial neural network (ANN) model is proposed to forecast the hotel occupancy in Puerto Rico. Neural networks is a nonparametric and data based technique that have been used in tourism demand forecasting for its flexibility and capability of mapping nonlinear complex functions. The forecast performance of artificial neural networks will be evaluated and compared to the performance of traditional Autoregressive Integrated Moving Average (ARIMA) time series models. Overnight stays is used as a measure of tourism demand. Monthly data from the Tourism Company of Puerto Rico from 2000 to 2014 is used.

Keywords: Neural networks, forecasting, hotel occupancy, Puerto Rico

Resumen

En este proyecto se propone un modelo de red neuronal artificial para predecir la ocupación hotelera en Puerto Rico. Las redes neuronales es una técnica no-paramétrica y basada en datos que se ha utilizado en la predicción de la demanda del sector turístico por su flexibilidad y capacidad de mapear funciones complejas no lineales. El desempeño de las redes neuronales artificiales se evaluará y se compara con el desempeño de los modelos tradicionales ARIMA de series de tiempo. La estancia en hotel se utiliza como una medida de la demanda turística. Se utilizan los datos mensuales de la Compañía de Turismo de Puerto Rico durante el periodo de 2000 al 2014.

Palabras claves: Redes neurales, predicción, ocupación hotelera, Puerto Rico

Introduction

Puerto Rico has been in an economic recession since 2006. Some economists argue that this prolonged recession can be assessed as an economic depression. The government has identified the tourism industry as a key player in helping the island pull out of the economic crisis. In 2014, the tourism industry accounted for close to 7% of the GDP, approximately 53,000 jobs, 13,500 of them in the hotel industry. In the period from January to June 2014, the hotel occupancy experienced an increment of 4% compared to the same period in 2013. In this particular context, it is critical to have appropriate techniques to forecast accurately the tourism demand in Puerto Rico.

Tourism has grown as an important sector of many economies. The importance to meet the sector's demands has drawn many researchers to study different techniques to forecast them. There are several measures to study tourism demands: tourist arrivals, tourist expenditures in the destination, tourism revenues, tourism employment, and overnight stays (Claveria & Torra, 2014). The most popular measure is tourist arrivals. Claveria & Torra work with tourist arrivals and overnight stays. Few papers have addressed the forecasting of tourism demands using overnight stays as a proxy to compare with tourist arrivals. This paper will use overnight stays as a measure of tourism demand.

Besides the measure used to study tourism demand, there are many statistical methods used for the forecasting of the measures. On the parametric side, time series models have been widely used in the forecasting process, particularly ARIMA models. Petrevska (2012), for example, identified an ARIMA (1,1,1) model to forecast Macedonia tourism demand measured by tourist arrivals.

Artificial Neural Networks (ANN) is a nonparametric and data based technique from the family of artificial intelligence methods that, although mostly applied in other fields, have also been used in tourism demand forecasting. This is a modelling alternative without the suppositions of the parametric counterpart. ANN models are capable of mapping linear or nonlinear functions without knowing beforehand the relationship between the input (independent) variables and the output (dependent) variables, introducing flexibility in the modeling process. As Chen et.al.

(2012) specifies: “There have been many studies using artificial neural networks (ANN) for tourism demand forecasting. These studies indicate a growing interest in using ANN as useful techniques for forecasting tourism demand, due to their ability to capture subtle functional relationships within the empirical data, even though the underlying relationships are unknown or hard to describe.”

This paper applies an ANN approach to forecast the hotel occupancy in Puerto Rico, following Claveria & Torra (2014). The forecast performance of these models will be evaluated and compared to the performance of traditional Autoregressive Integrated Moving Average (ARIMA) time series models. Monthly data from the Tourism Company of Puerto Rico from 2000 to 2014 will be used.

Methodology

The data set consists of monthly data of tourist overnight stays from foreign countries to Puerto Rico from 2000 to 2014 collected by the Tourism Company of Puerto Rico. After cleaning the data, a data set was created that includes monthly data, in the period of 2000-2014, of the following variables: total accommodation registration, non-resident registration, resident registration, occupancy rates, room nights rented, room nights available, number of guests, and the average daily rate. Overnight stays, represented by accommodation registrations, is used as a measure of tourism demand.

ARIMA(p,d,q) models are widely used to model time series data. It is a combination of an autoregressive model AR(p) and a moving average model MA(q). If the data requires differencing to achieve stationarity, the d stands for the number of differences taken. These models are compared to the neural network models in their forecast ability to predict hotel registrations.

For the ANN models, the methodology used by Claveria &Torra (2014) is followed. They use the multi-layer perceptron (MLP) method, one of the most popular neural network models used in time series. As Wei and Chen (2012) explain: “The architecture of MLP consists of multiple layers, which include an input layer, one or more hidden layers, and an output layer. Each layer comprises several neurons connected to the neurons in neighboring layers. Since MLP contains many interacting nonlinear neurons in multiple layers, it can capture complex phenomena."

The MLP specification used by Claveria &Torra (2014) is used:

$$x_t = f\left(\beta_0 + \sum_{j=1}^{q} \beta_j g(x_{t-1}\varphi_{ij} + \varphi_{0j})\right),$$
$$\{\varphi_{ij}, i = 0,1,\ldots, p, \; j = 1,\ldots, q\}$$
$$\{\beta_j \, , \; j = 0,1,\ldots, q\}$$

where f is the output function, g is the activation function, p is the number of inputs (lagged values), q is the number of neurons or nodes in the hidden layer, x_t is the output (Registrations at time t), x_{t-1} (Registrations at time t-1) is the input, β_j are the weights connecting the output with the hidden layer and φ_{ij} are the weights connecting the input with the hidden layer. This paper considers one hidden layer in a multilayer feed-forward network, where each node in a layer receives inputs from the prior layer. The inputs are combined through a weighted linear combination in each node, and then modified by a nonlinear function. The output is then the input to the next layer. Several combinations of lagged inputs (p) and number of neurons are considered.

The data set is divided in three smaller sets to create a training data set, a validation data set, and a test data set. The first 50% observations are on the training set, 40% on the validation set and the last 10% of the observations in the test data set. The statistical analysis are done with R. The models are compared using the root mean squared forecast error (RMSFE).

Empirical Results

This section presents the descriptive and inferential analyses to forecast hotel registration in Puerto Rico from 2000-2014. In Figure 1 it can be seen that hotel registrations have been in an upward trend, with a seasonal behavior. Figure 2 presents the hotel registrations for non-residents and residents. The resident registrations have increased at a faster pace than the non-residents.

Figure 1. Monthly Accommodations Registrations in Puerto Rico, 2000-2014

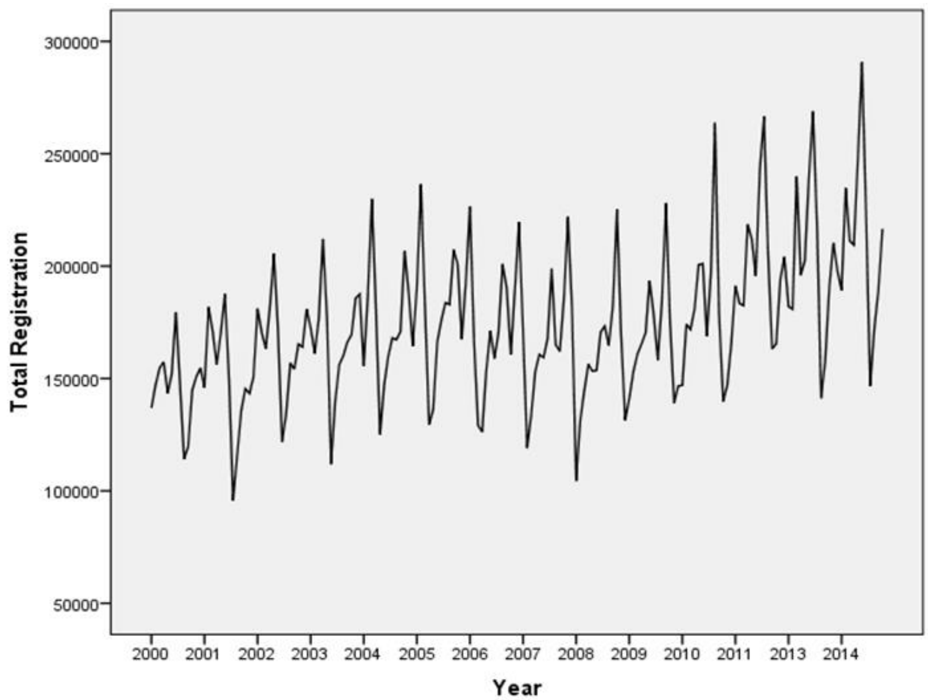


Figure 2. Monthly Accommodations Registrations in Puerto Rico, Total, Non Residents and Residents, 2000-2014

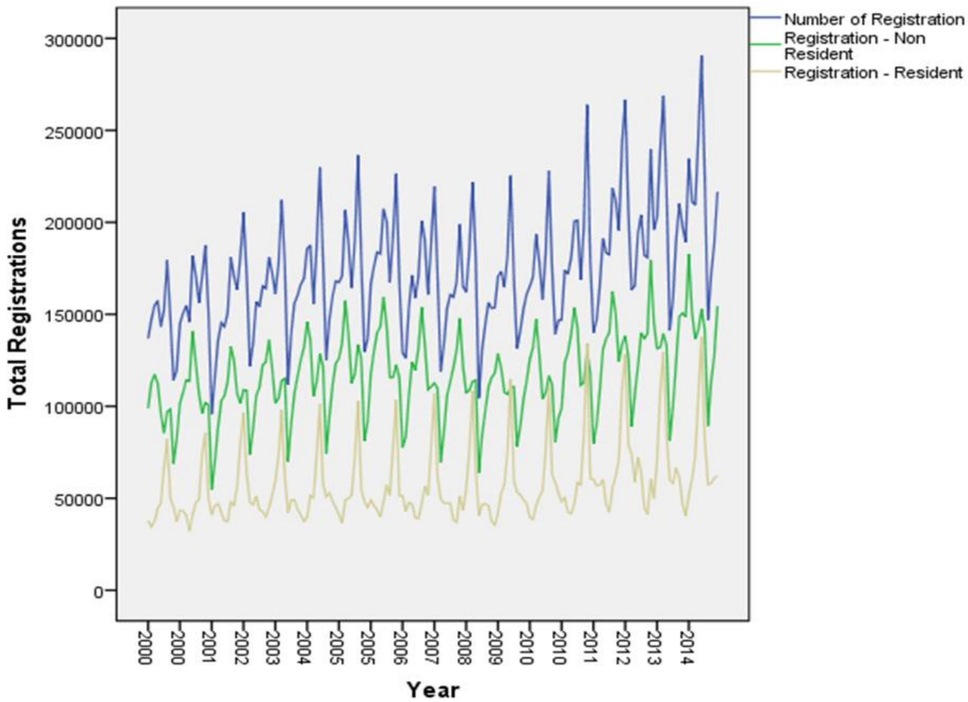


Figure 3. Decomposition of Monthly Accommodations Registrations

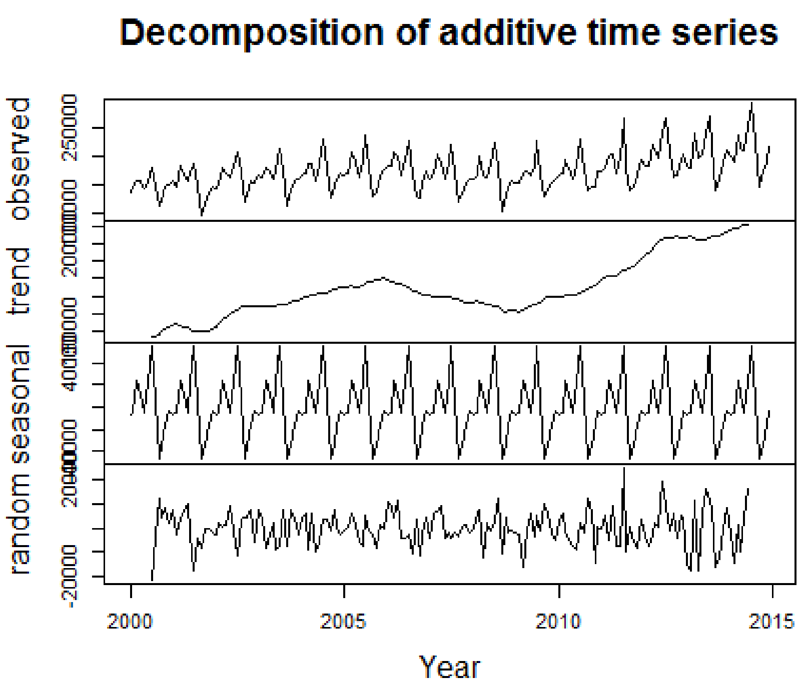


Table 1 Accuracy measures for model comparison to forecast total registration

Model	Mean absolute percentage error (MAPE)	Mean absolute deviation (MAD)	Mean squared deviation (MSD)	Root Mean Squared Forecast Error (RMSFE)
Linear Trend Model	13	21936	826787115	
Quadratic Trend Model	13	21749	815633972	
S-Curve Trend Model	13	21654	827847748	
Aditive Trend Model with Seasonal Component	6	10206	155300322	
Multiplicative Model with seasonal Component	6	9800	144270630	

Table 2 Comparison of ARIMA and ANN models

Model	AIC_c	Log-likelihood		
ARIMA(2,1,2) with drift	4,101.86	2044.69		
ARIMA(2,0,2)	-214.59	113.54		
$ARIMA(2,1,2)(0,1,1)_{12}$	19.33	-1764.22		
$ARIMA(2,1,2)(1,1,0)_{12}$	19.39	-1768.63		

References

Chen, C. F., Lai, M. C., & Yeh, C. C. (2012). Forecasting tourism demand based on empirical mode decomposition and neural network. *Knowledge-Based Systems*, *26*, 281-287.

Claveria, O., & Torra, S. (2014). Forecasting tourism demand to Catalonia: Neural networks vs. time series models. *Economic Modelling*, *36*, 220-228.

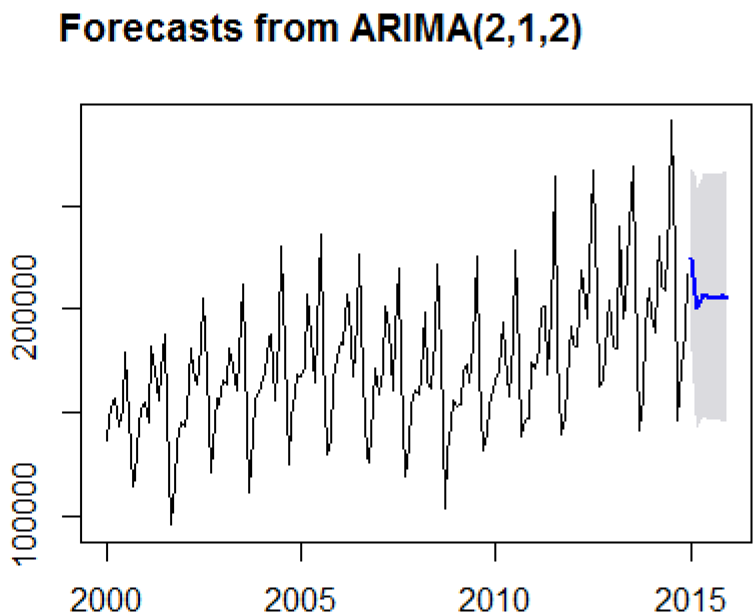
Cho, V. (2001). Tourism forecasting and its relationship with leading economic indicators. *Journal of Hospitality & Tourism Research*, *25*(4), 399-420.

Huarng, K. H., Hui-Kuang Yu, T., Moutinho, L., & Wang, Y. C. (2012). Forecasting tourism demand by fuzzy time series models. *International Journal of Culture, Tourism and Hospitality Research*, *6*(4), 377-388.

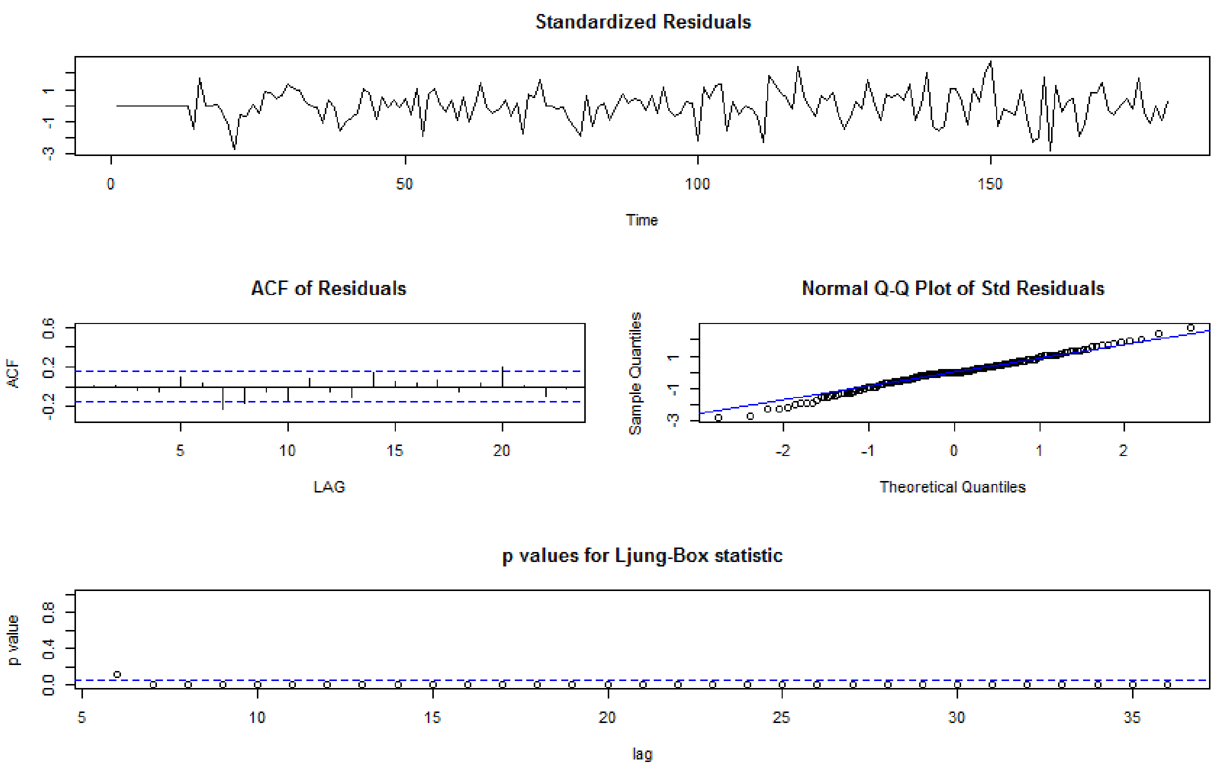
Petrevska, B. (2012). Forecasting international tourism demand: The evidence of Macedonia. *UTMS Journal of Economics*, *3*(1), 45-55.

Song, H., & Li, G. (2008). Tourism demand modelling and forecasting—A review of recent research. *Tourism Management*, *29*(2), 203-220.

Wei, Y., & Chen, M. C. (2012). Forecasting the short-term metro passenger flow with empirical mode decomposition and neural networks. *Transportation Research Part C: Emerging Technologies*, *21*(1), 148-162.

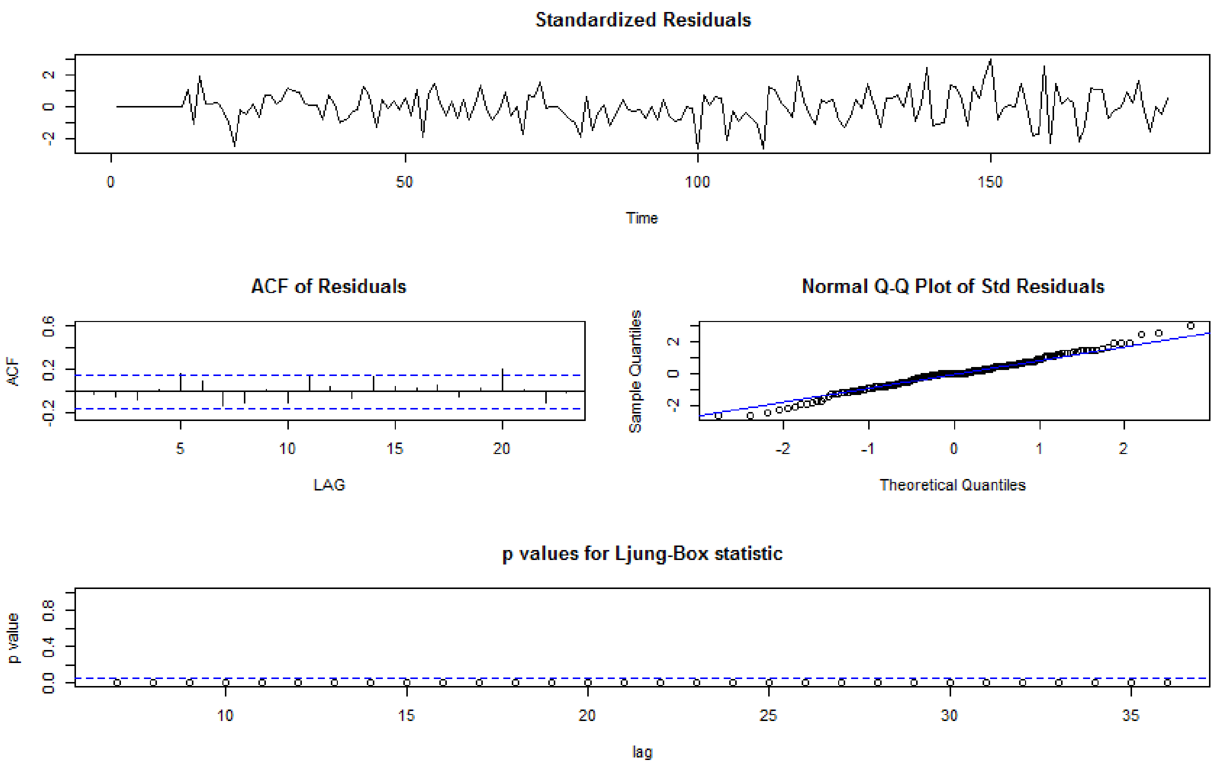

Forecasts from ARIMA(2,1,2)
100000
200000
2000
2005
2010
2015

```
> sarima(Registrations, 2,1,2,1,1,0,12)
$fit

Call:
stats::arima(x = xdata, order = c(p, d, q), seasonal = list(order = c(P, D,
    Q), period = S), include.mean = !no.constant, optim.control = list(trace
= trc,
    REPORT = 1, reltol = tol))

Coefficients:
          ar1      ar2     ma1      ma2     sar1
      -1.1081  -0.2552  0.4297  -0.5411  -0.3085
s.e.   0.1181   0.1179  0.1131   0.1133   0.0773

sigma^2 estimated as 90685497:  log likelihood = -1768.63,  aic = 3549.25

$AIC
[1] 19.37846

$AICc
[1] 19.39227

$BIC
[1] 18.46716
```



```
> fit <- auto.arima(Registrations, lambda=0, d=0, D=1, max.order=9,
+                   stepwise=FALSE, approximation=FALSE)
>
> tsdisplay(residuals(fit))
> fit
Series: Registrations
ARIMA(4,0,1) with non-zero mean
Box Cox transformation: lambda= 0

Coefficients:
          ar1     ar2      ar3     ar4     ma1  intercept
      -0.0003  0.1661  -0.1473  0.3288  0.9723    12.0519
s.e.   0.0833  0.0838   0.0818  0.0792  0.0396     0.0282

sigma^2 estimated as 0.01605:  log likelihood=115.55
AIC=-217.1   AICc=-216.45   BIC=-194.75
```

```
> sarima(Registrations,4,0,1,0,1,1,12)
$fit

Call:
stats::arima(x = xdata, order = c(p, d, q), seasonal = list(order = c(P, D,
    Q), period = S), xreg = constant, optim.control = list(trace = trc, REPOR
T = 1,
    reltol = tol))

Coefficients:
          ar1     ar2     ar3     ar4     ma1     sma1  constant
      -0.6560  0.5075  0.5114  0.1591  0.9717  -0.4282  379.2952
s.e.   0.0833  0.0831  0.0821  0.0809  0.0303   0.0993  141.2223

sigma^2 estimated as 83098990:  log likelihood = -1772.44,  aic = 3560.87

$AIC
[1] 19.31332

$AICc
[1] 19.32911

$BIC
[1] 18.43749

Training set error measures:
                   ME     RMSE      MAE       MPE     MAPE     MASE         AC
F1
Training set 1131.622 8929.612 6736.656 0.2959299 3.930643 0.284672 -0.038470
05
```

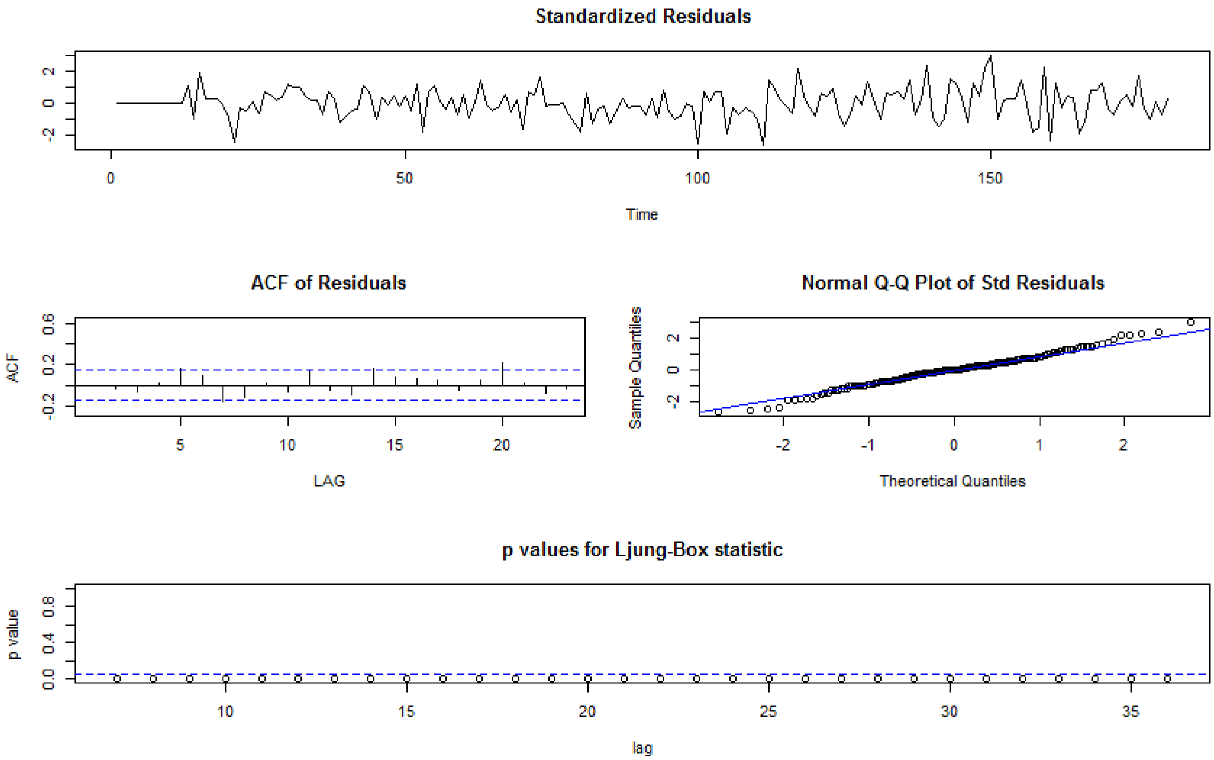


```
> sarima(Registrations,4,0,1,1,1,0,12)
```


```
> sarima(Registrations,4,0,1,1,1,0,12)
```

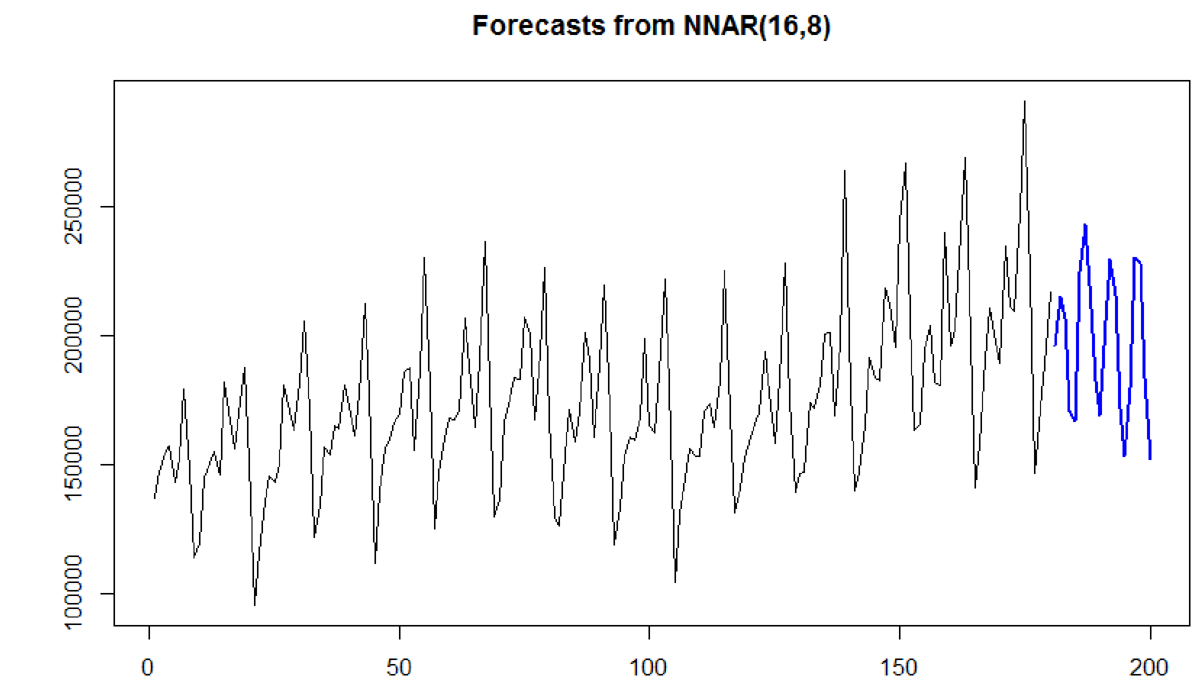


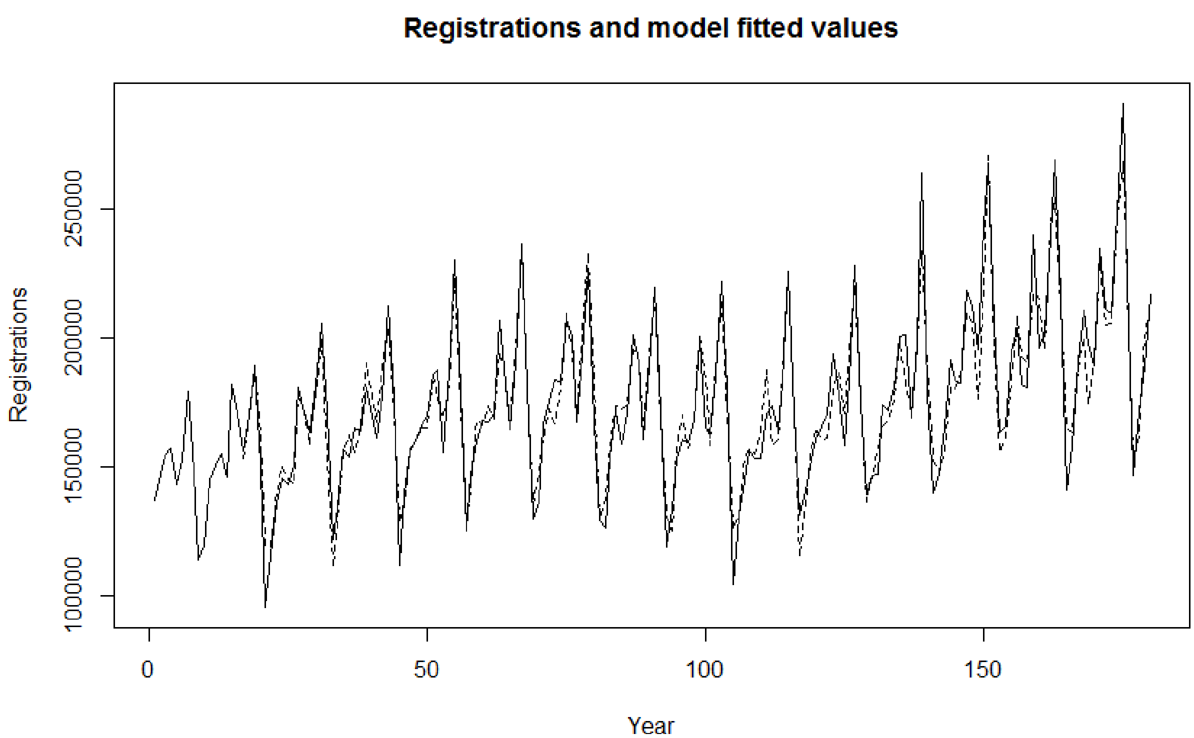
```
$fit

Call:
stats::arima(x = xdata, order = c(p, d, q), seasonal = list(order = c(P, D,
    Q), period = S), xreg = constant, optim.control = list(trace = trc, REPOR
T = 1,
    reltol = tol))

Coefficients:
          ar1     ar2     ar3     ar4     ma1     sar1  constant
      -0.6895  0.4752  0.4979  0.1499  0.9825  -0.2785  387.9776
s.e.   0.0792  0.0843  0.0848  0.0808  0.0255   0.0785  161.9249

sigma^2 estimated as 86660787:  log likelihood = -1775.26,  aic = 3566.53

$AIC
[1] 19.35529

$AICc
[1] 19.37108

$BIC
[1] 18.47946

Training set error measures:
                   ME     RMSE      MAE       MPE    MAPE      MASE         AC
F1
Training set 1121.667 9100.023 6879.375 0.3280551 4.02482 0.2907029 -0.033196
04
```

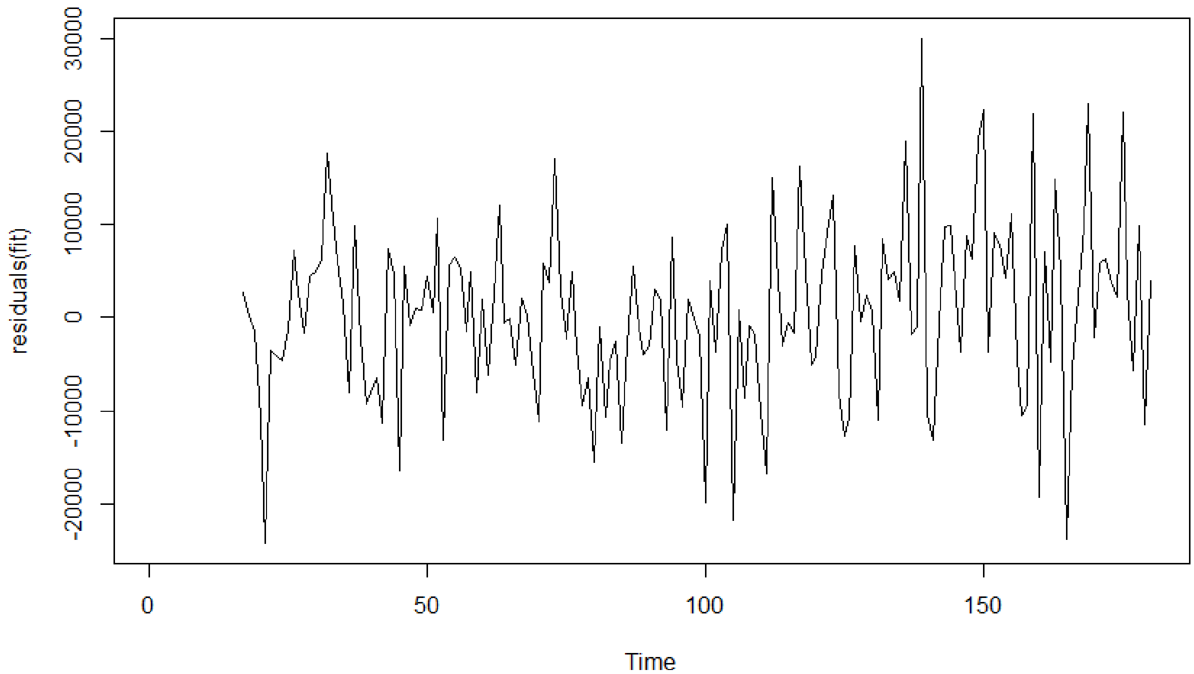

Standardized Residuals
Time
ACF of Residuals
ACF
LAG
Normal Q-Q Plot of Std Residuals
Sample Quantiles
Theoretical Quantiles
p values for Ljung-Box statistic
p value
lag

Neural:

```
Series: Registrations
Model:  NNAR(16,8)
Call:   nnetar(x = Registrations, lambda = 0)

Average of 20 networks, each of which is
a 16-8-1 network with 145 weights
options were - linear output units

sigma^2 estimated as 86546446
> plot(forecast(fit,h=20))
 > nnfit<-forecast(fit, h=12)
> nnfit
    Point Forecast
181       196001.4
182       215132.3
183       206674.7
184       170442.9
185       166808.9
186       224341.1
187       243054.5
188       219304.0
189       182414.4
190       169125.6
191       205697.9
192       229718.1

ME      RMSE       MAE          MPE      MAPE      MASE        ACF1
Training set 498.8819 9303.034 7165.718 -0.1753113 4.216084 0.302803 0.048332
75
```

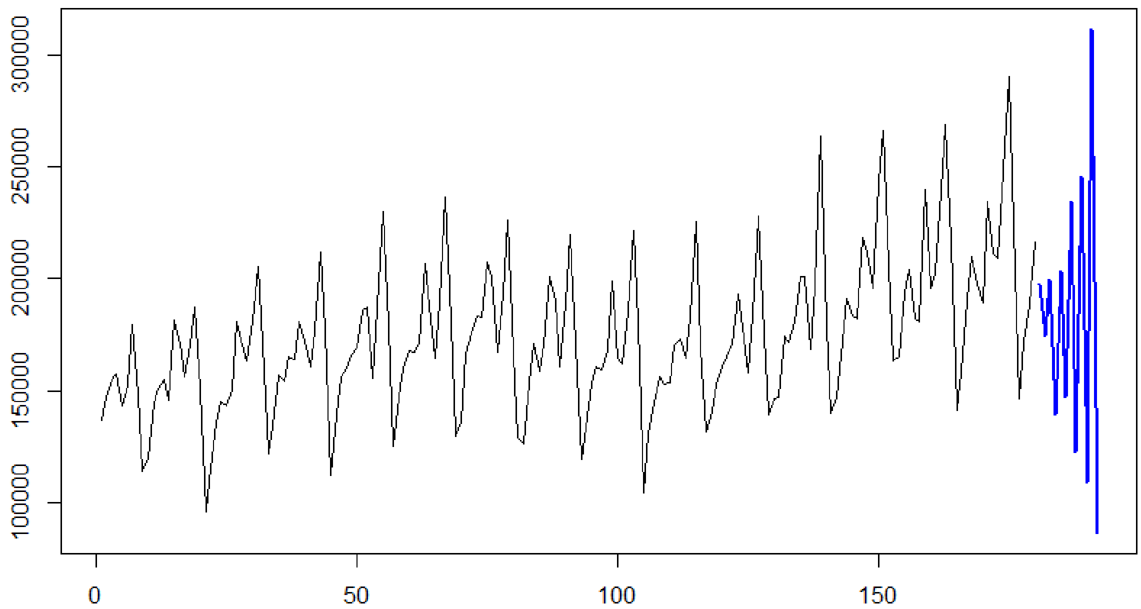

Forecasts from NNAR(16,8)
100000
150000
200000
250000
0
50
100
150
200


Registrations and model fitted values
Registrations
100000
150000
200000
250000
0
50
100
150
Year


residuals(fit)
-20000
-10000
0
10000
20000
30000
0
50
100
150
Time

```
> mod1<-nnetar(Registrations, 4,P=1,lambda=0)
> mod1
Series: Registrations
Model:  NNAR(4,2)
Call:   nnetar(x = Registrations, p = 4, P = 1, lambda = 0)

Average of 20 networks, each of which is
a 4-2-1 network with 13 weights
options were - linear output units

sigma^2 estimated as 533142561
> mod1fit<-forecast(mod1, h=12)
> mod1fit
    Point Forecast
181      197508.95
182      174460.30
183      199362.00
184      139660.80
185      203457.32
186      146849.66
187      234766.67
188      122617.64
189      245837.03
190      109125.62
191      311708.32
192       86457.74
> summary(mod1)
          Length Class        Mode
x         180    ts           numeric
m           1    -none-       numeric
p           1    -none-       numeric
P           1    -none-       numeric
scale       1    -none-       numeric
size        1    -none-       numeric
lambda      1    -none-       numeric
model      20    nnetarmodels list
fitted    180    ts           numeric
residuals 180    ts           numeric
lags        4    -none-       numeric
series      1    -none-       character
method      1    -none-       character
call        5    -none-       call


ME      RMSE      MAE        MPE      MAPE      MASE      ACF1
Training set 1713.277 23089.88 18099.48 -0.9114618 10.45446 0.7648327 0.10213
19
```

Forecasts from NNAR(4,2)



Registrations and model fitted values

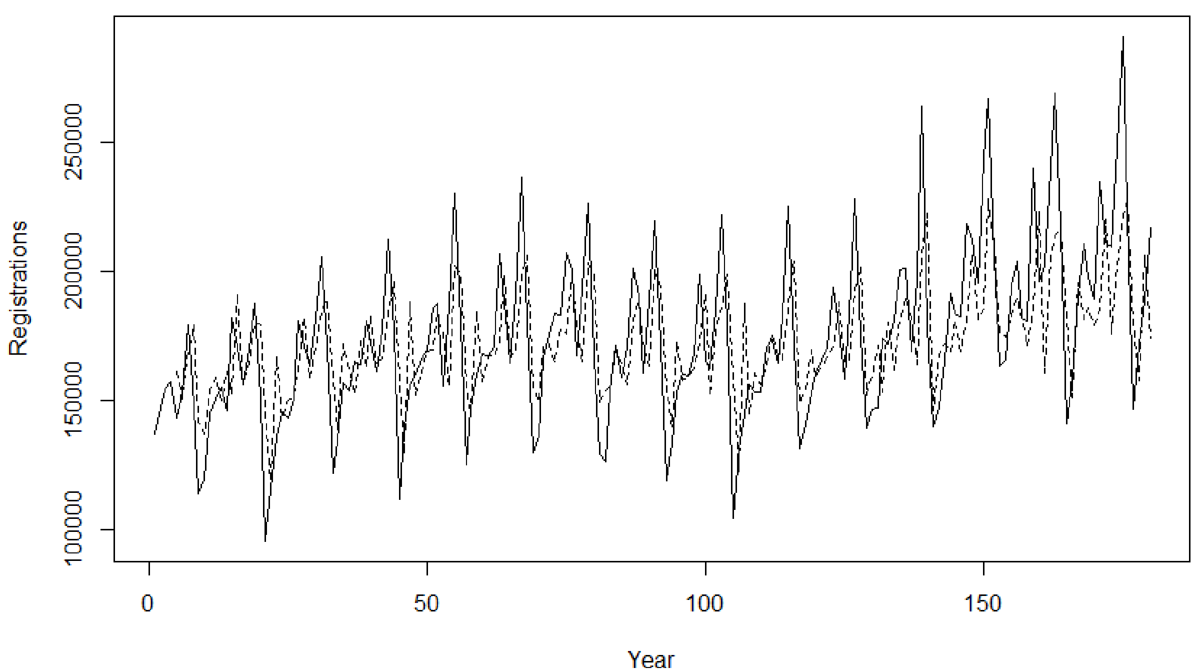


Figure 3 Monthly Room Nights Rented and Room Nights Available in Puerto Rico, 2000-2014

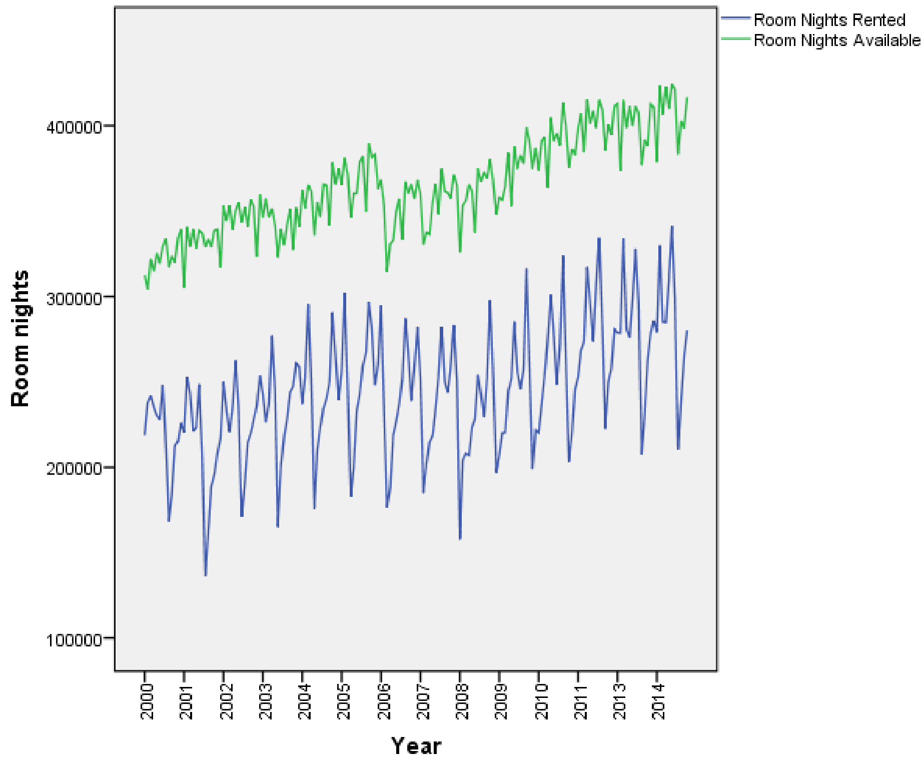


Figure 4 Accommodations Occupancy Rate (%) in Puerto Rico, 2000-2014

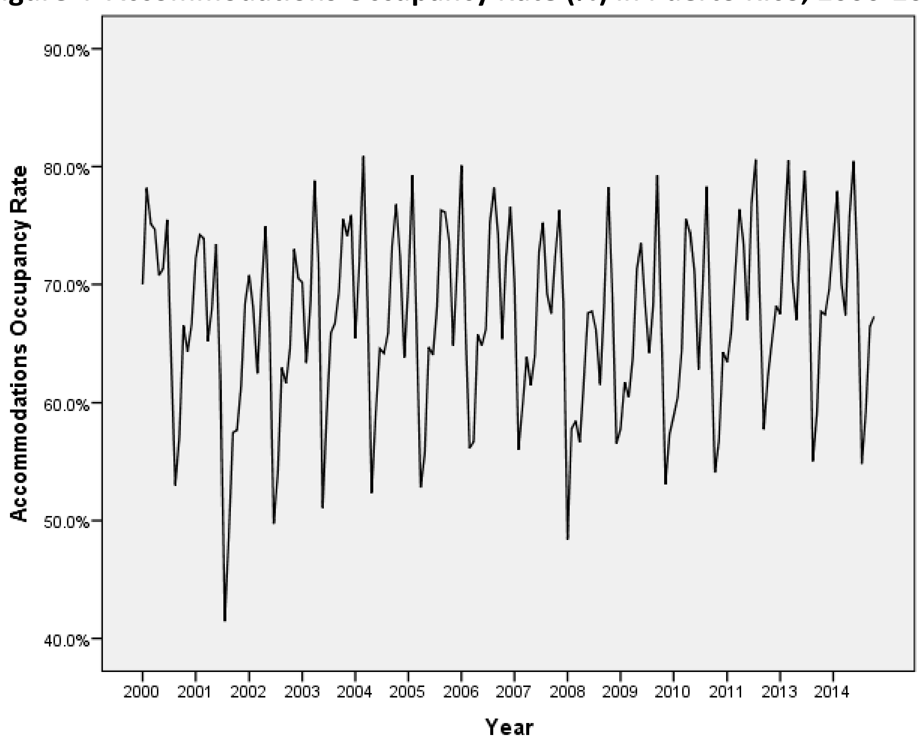
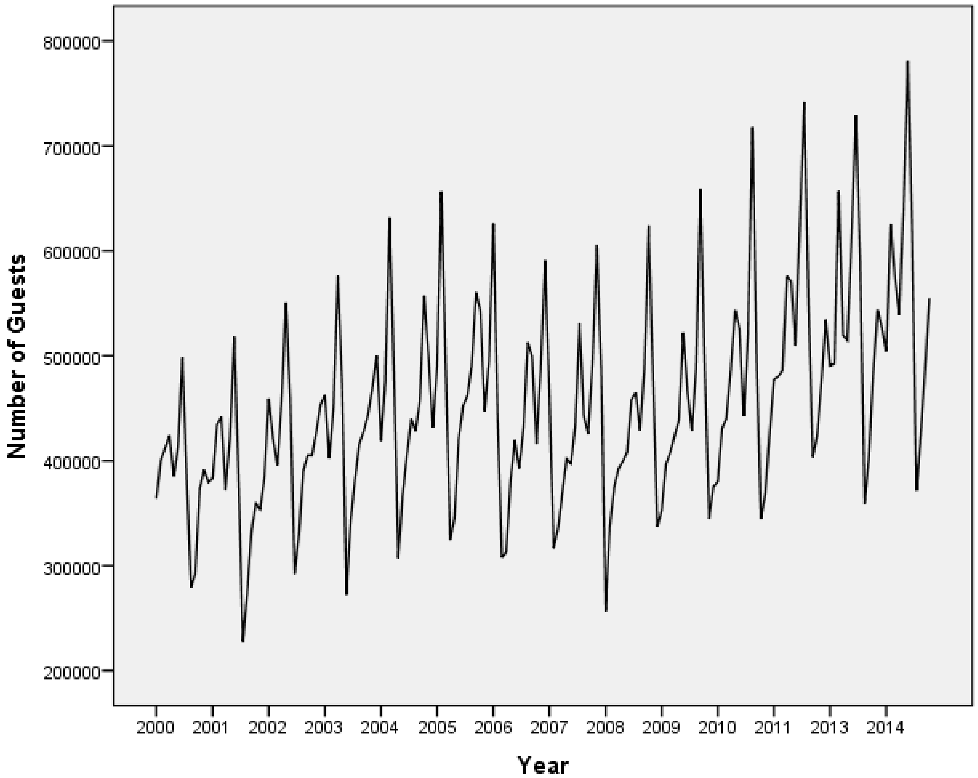


Figure 5 Monthly Number of Guests in Puerto Rico, 2000-2014



Measures of accuracy (time series analysis)

Use these statistics to compare the fits of different forecasting and smoothing methods. Minitab computes three measures of accuracy of the fitted model: MAPE, MAD, and MSD. The three measures are not very informative by themselves, but you can use them to compare the fits obtained by using different methods. For all three measures, smaller values generally indicate a better fitting model.

- **Mean absolute percentage error (MAPE)** – Expresses accuracy as a percentage of the error. Because this number is a percentage, it may be easier to understand than the other statistics. For example, if the MAPE is 5, on average the forecast is off by 5%.

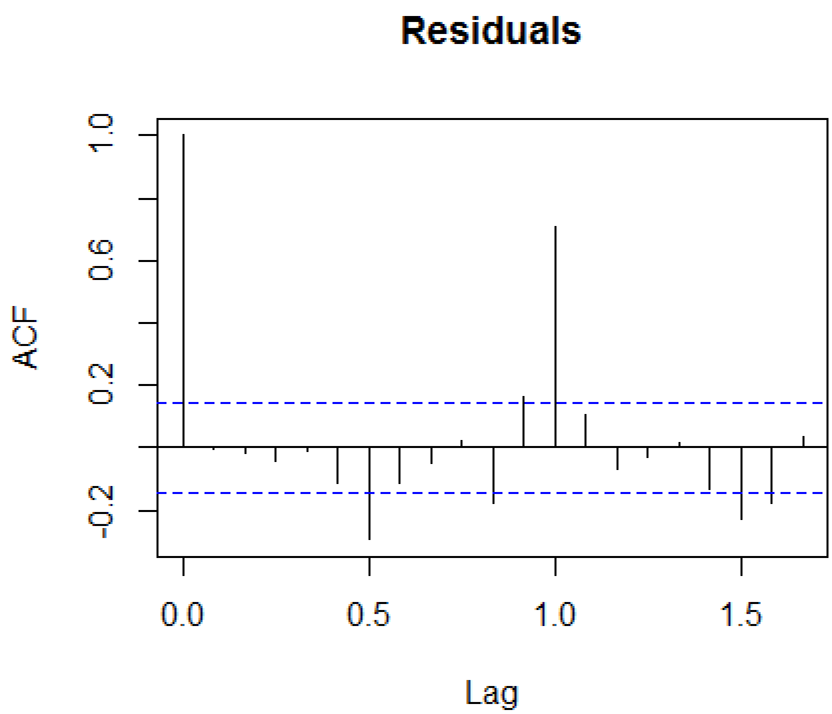
- **Mean absolute deviation (MAD)** – Expresses accuracy in the same units as the data, which helps conceptualize the amount of error. Outliers have less of an affect on MAD than on MSD.

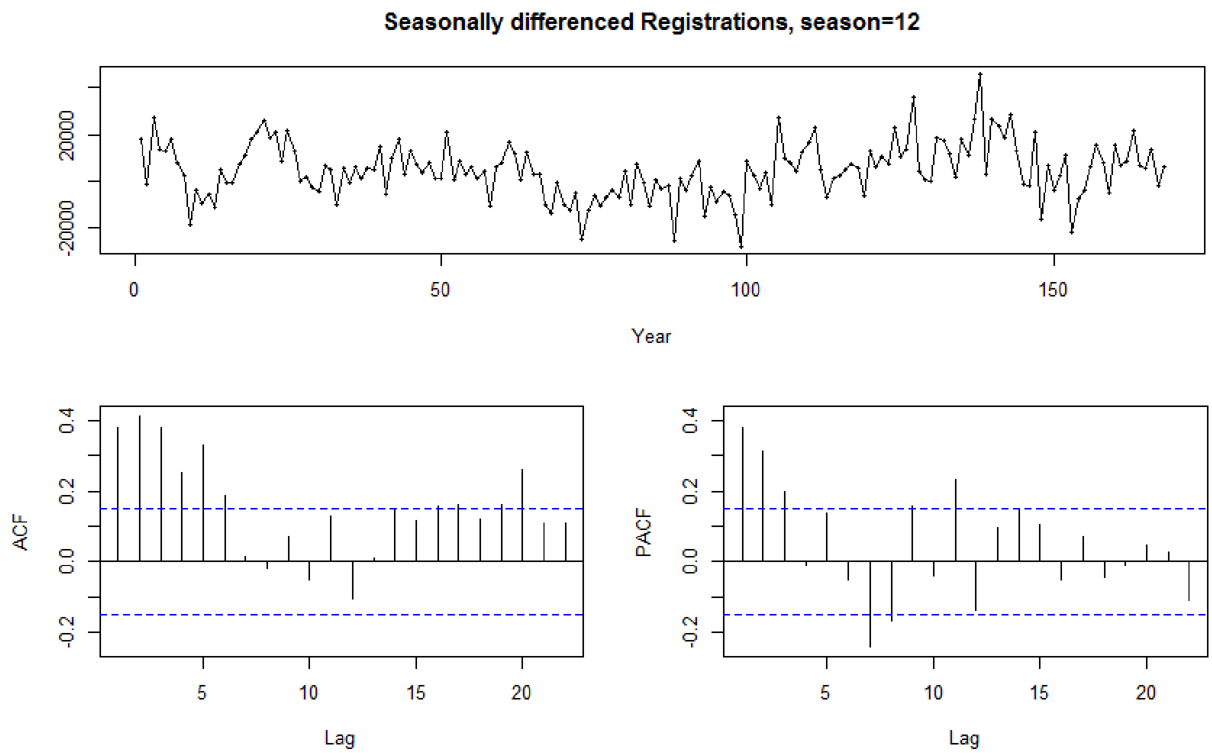
- **Mean squared deviation (MSD)** – A commonly-used measure of accuracy of fitted time series values. Outliers have more influence on MSD than MAD.

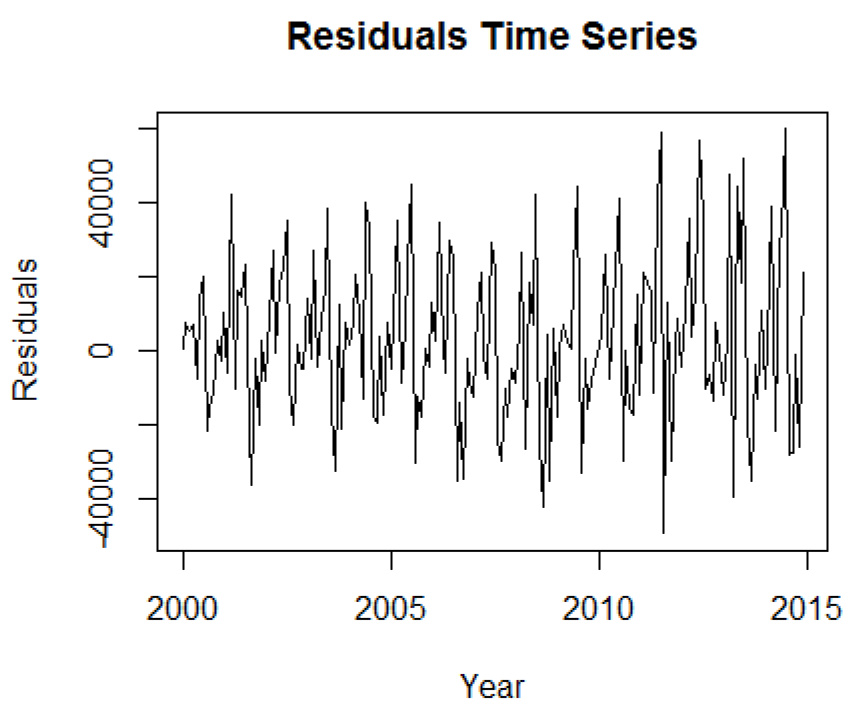
Concluding Remarks

Appendix

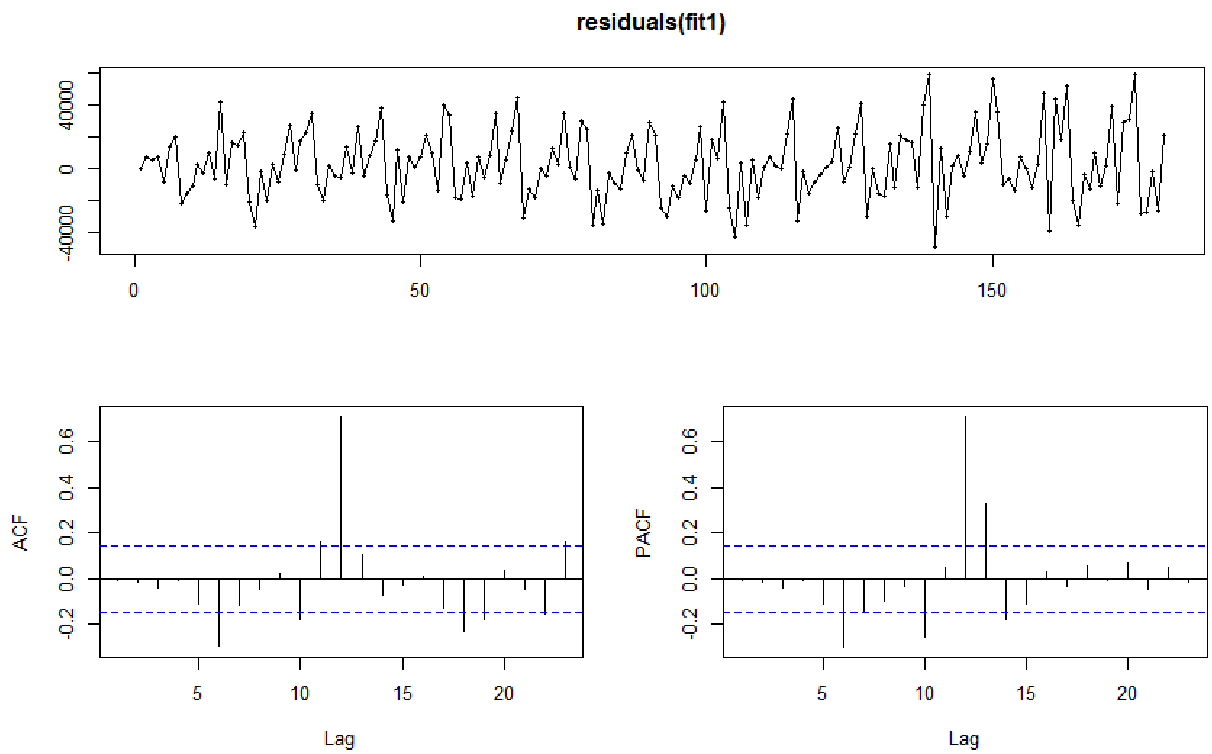
```
Series: Registrations
ARIMA(2,1,2) with drift

Coefficients:
         ar1      ar2      ma1      ma2    drift
      0.2323  -0.2853  -0.3848  -0.5416  338.211
s.e.  0.1110   0.0958   0.1040   0.1041  132.699

sigma^2 estimated as 481927908:  log likelihood=-2044.69
AIC=4101.38   AICc=4101.86   BIC=4120.5


auto.arima(Registrations, ic="bic")
Series: Registrations
ARIMA(2,1,2)

Coefficients:
         ar1      ar2      ma1      ma2
      0.2304  -0.2914  -0.3577  -0.5195
s.e.  0.1111   0.0948   0.1045   0.1005

sigma^2 estimated as 494404686:  log likelihood=-2046.74
AIC=4103.47   AICc=4103.82   BIC=4119.41



Call:
arima(x = Regts, order = c(2, 1, 2))

Coefficients:
         ar1      ar2      ma1      ma2
      0.2304  -0.2914  -0.3577  -0.5195
s.e.  0.1111   0.0948   0.1045   0.1005

sigma^2 estimated as 494404686:  log likelihood = -2046.74,  aic = 4103.47

> forecast.Arima(regsarima,h=12, level=c(95))
         Point Forecast    Lo 95    Hi 95
Jan 2015       224018.7 180438.5 267598.9
Feb 2015       206695.6 148853.3 264537.9
Mar 2015       200540.2 142680.6 258399.8
Apr 2015       204169.9 146058.4 262281.4
May 2015       206799.7 148570.9 265028.5
Jun 2015       206347.9 147601.9 265093.8
Jul 2015       205477.5 146419.1 264535.9
Aug 2015       205408.6 146180.3 264637.0
Sep 2015       205646.4 146238.4 265054.4
Oct 2015       205721.2 146093.9 265348.5
Nov 2015       205669.2 145817.0 265521.4
Dec 2015       205635.4 145569.9 265700.9
```


Residuals
ACF
1.0
0.6
0.2
-0.2
0.0
0.5
1.0
1.5
Lag

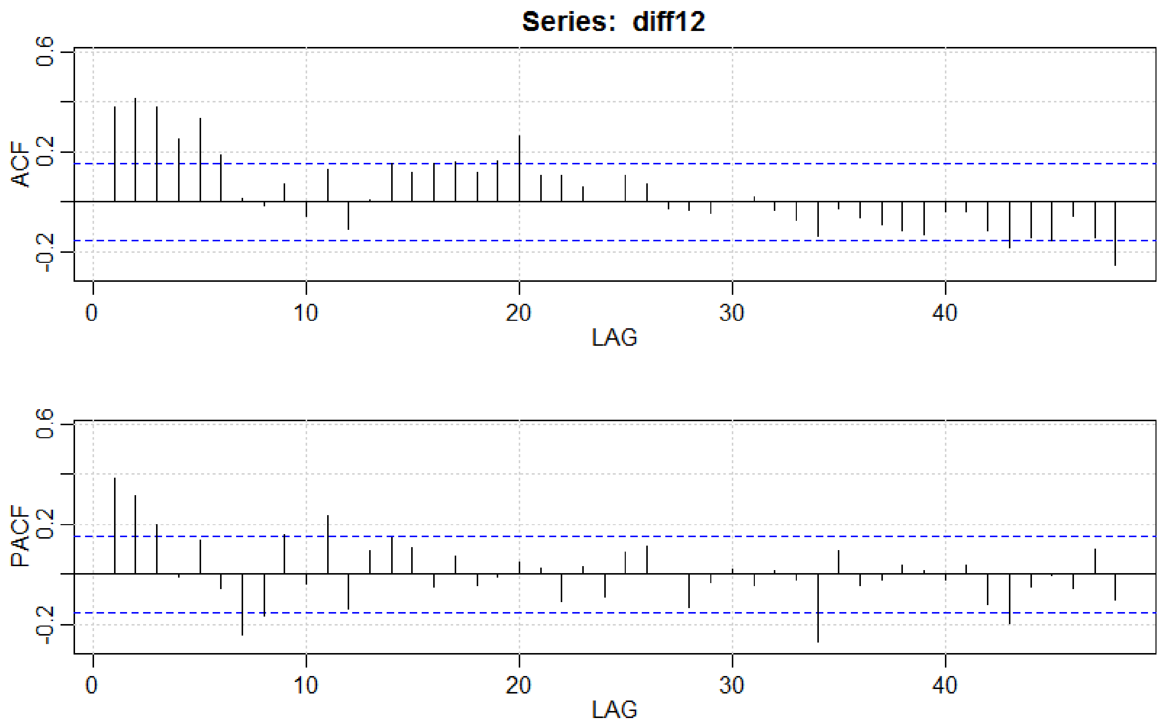

Seasonally differenced Registrations, season=12
20000
-20000
0
50
100
150
Year
ACF
0.4
0.2
0.0
-0.2
5
10
15
20
Lag
PACF
0.4
0.2
0.0
-0.2
5
10
15
20
Lag


Residuals Time Series
Residuals
40000
0
-40000
2000
2005
2010
2015
Year

```
Box-Ljung test

data:  regs.forecast$residuals
X-squared = 155.56, df = 20, p-value < 2.2e-16


 > fit1 <- Arima(Registrations, order=c(2,1,2), seasonal=c(0,1,1))
>
> tsdisplay(residuals(fit1))
> fit1
Series: Registrations
ARIMA(2,1,2)

Coefficients:
         ar1      ar2      ma1      ma2
      0.2304  -0.2914  -0.3577  -0.5195
s.e.  0.1111   0.0948   0.1045   0.1005

sigma^2 estimated as 494404686:  log likelihood=-2046.74
AIC=4103.47   AICc=4103.82   BIC=4119.41
```



```
> fit <- Arima(Registrations, order=c(2,0,2), seasonal=c(0,1,1), lambda=0)
> tsdisplay(residuals(fit))
> fit
Series: Registrations
ARIMA(2,0,2) with non-zero mean
Box Cox transformation: lambda= 0

Coefficients:
         ar1      ar2      ma1      ma2  intercept
      1.0258  -0.0345  -0.1252  -0.7636    12.0597
s.e.  0.0995   0.0980   0.0684   0.0594     0.0857

sigma^2 estimated as 0.0164:  log likelihood=113.54
AIC=-215.07   AICc=-214.59   BIC=-195.91
```



```
> sarima(Registrations, 2,1,2,0,1,1,12)
$fit

Call:
stats::arima(x = xdata, order = c(p, d, q), seasonal = list(order = c(P, D, 
    Q), period = S), include.mean = !no.constant, optim.control = list(trace 
= trc, 
    REPORT = 1, reltol = tol))

Coefficients:
          ar1      ar2     ma1      ma2     sma1
      -1.0443  -0.2002  0.3725  -0.5735  -0.5057
s.e.   0.1296   0.1287  0.1202   0.1179   0.0959

sigma^2 estimated as 84881961:  log likelihood = -1764.22,  aic = 3540.43

$AIC
[1] 19.31233

$AICc
[1] 19.32614

$BIC
[1] 18.40102
```