Forecasting

EPK

06/02/2022

# Part 1

## Introduction

Forecasting is a strategy that uses previous data as inputs to make well-informed predictions about the direction of future trends. Forecasting is used by businesses to determine how to allocate their budgets or plan for anticipated expenses in future (Tuovila, n.d.). Since 1959 through 2021, US-seasonally adjusted personal consumption expenditure data is available for analysis and forecasting to determine expenditure in coming months using historical data. The given historical data is sequence of observation across time specified in years hence it follows a time series format for further development.

Pre-processing the data to discover missing values and imputing using appropriate method with time - series imputation. Forecasting methods such as simple forecasting, exponential and arima models, is the primary analysis to compare predictive capacity. These models gained a better understanding by analysing the accuracy of each model’s performance and determining the optimal model with fewest errors.

## Model Analysis

Installing and invoking the essential libraries for forecasting modeling is the first step in the analysis. Read the pce dataset into a data frame after adding the relevant libraries to do further pre-processing of missing data and discovering the structure of variables to check whether they are the correct class of variable if they do not need to be converted to the appropriate class type.

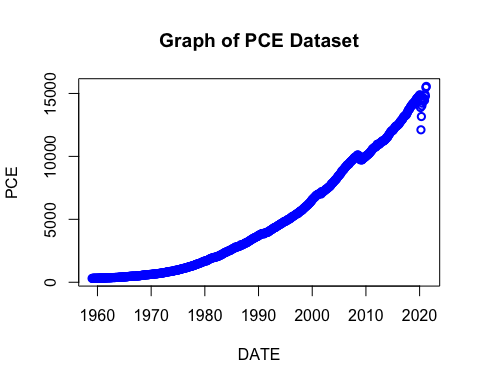
## 'data.frame': 748 obs. of 2 variables:  
## $ DATE: Factor w/ 748 levels "01/01/1959","01/01/1960",..: 1 64 127 190 253 315 377 439 501 563 ...  
## $ PCE : num 306 310 313 312 316 ...

The variable date is in factor format, which needs to be converted to the appropriate format, which is date, because the structure of the data must be accurate in order to perform multiple imputation.

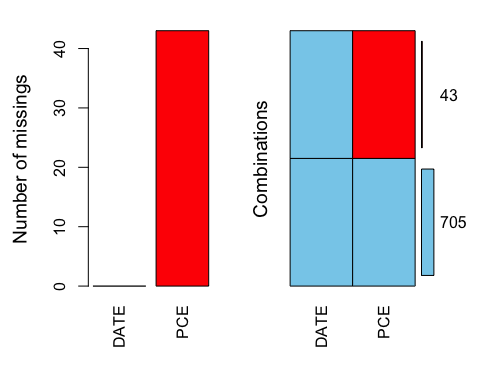
### Multiple Imputation

Imputation is a technique mainly used for replacing missing data in dataset. The technique used to impute missing values is multiple imputation where the process of substituting multiple values for each missing cell based on shared predictive distribution(Alice,2015).

## [1] "Number of incomplete Rows:43"



The graph shows that there are a few missing values, which are represented by unshaded circles.



##   
## iter imp variable  
## 1 1 PCE  
## 1 2 PCE  
## 1 3 PCE  
## 1 4 PCE  
## 1 5 PCE  
## 2 1 PCE  
## 2 2 PCE  
## 2 3 PCE  
## 2 4 PCE  
## 2 5 PCE  
## 3 1 PCE  
## 3 2 PCE  
## 3 3 PCE  
## 3 4 PCE  
## 3 5 PCE  
## 4 1 PCE  
## 4 2 PCE  
## 4 3 PCE  
## 4 4 PCE  
## 4 5 PCE  
## 5 1 PCE  
## 5 2 PCE  
## 5 3 PCE  
## 5 4 PCE  
## 5 5 PCE  
## 6 1 PCE  
## 6 2 PCE  
## 6 3 PCE  
## 6 4 PCE  
## 6 5 PCE  
## 7 1 PCE  
## 7 2 PCE  
## 7 3 PCE  
## 7 4 PCE  
## 7 5 PCE  
## 8 1 PCE  
## 8 2 PCE  
## 8 3 PCE  
## 8 4 PCE  
## 8 5 PCE  
## 9 1 PCE  
## 9 2 PCE  
## 9 3 PCE  
## 9 4 PCE  
## 9 5 PCE  
## 10 1 PCE  
## 10 2 PCE  
## 10 3 PCE  
## 10 4 PCE  
## 10 5 PCE

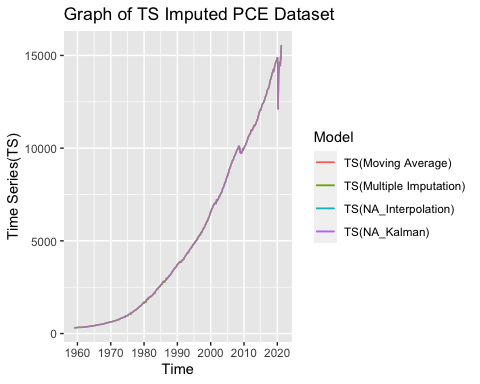
## [1] "Number of incomplete Rows:0"

The shared predictive distribution method replaces all NA values with appropriate personal consumption values.

### Analysis of Time Series Imputation

Here, dataset is converted to time series model using the function ‘ts’ where frequency denotes number of periods, h=12 denotes it is monthly dataset.

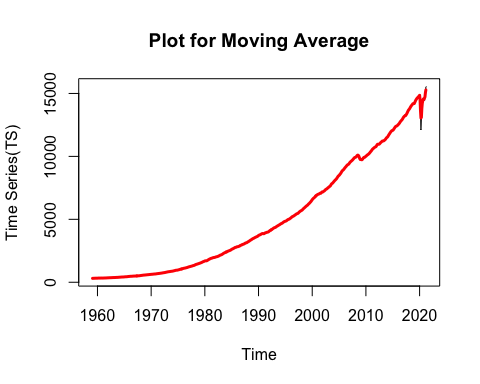
The dataset obtained using time series imputed approach is compared with dataset obtained using multiple imputation. The ‘na kalman’ time series imputation has the smallest difference of all the others and is used for further investigation.



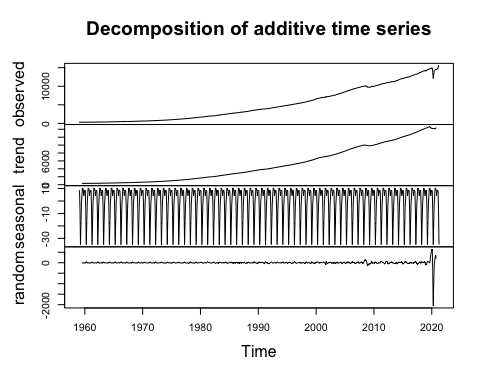
The graph clearly illustrates time series modeled using various imputation methods, and it is evident that the error between all four imputation methods is very low. In addition, the graph demonstrates that personal consumption expenditures are on the rising trend over time.

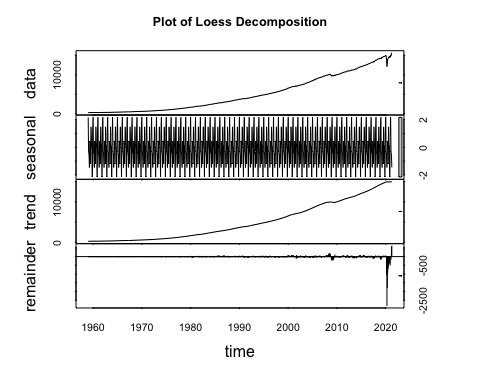
### Moving Average

Moving average is a smoothing technique that helps to understand the trend in the time series by smoothening the fluctuation during COVID-19 period.

 ### Decomposition

Decomposition is used to extract the trend and seasonality from the observed time series dataset, and the type is additive because there is no significant seasonality fluctuation over time for the given dataset.

 The fundamental disadvantage of normal additive decomposition is that it ignores irregular components and assumes seasonal components over time because it does not use the initial and last few observations, which are NA values. In order to address this challenge, Loess decomposition makes use of loess functionality to help with trend and can handle any form of seasonality over time.



### 1

### Forecast Modeling

To model time series and assess the accuracy of each model, divide the dataset into 80-20 percent training and test datasets, which will help you identify the optimal model with the minimal performance errors.

### Simple Forecasting Methods

Some simple forecasting approaches include training the model with training data and forecasting future data by passing the number of periods equal to the number of observations in test data to assess the model’s performance.

The forecasting techniques are trained using training data to determine the model’s performance, as shown below, and a graphical depiction of the forecast generated with original time series data is used to evaluate the various methods.

## [1] "Averge Method(meanf)"

## ME RMSE MAE MPE MAPE MASE  
## Training set 4.986869e-14 2879.078 2439.910 -192.11520 226.27118 12.30834  
## Test set 8.929285e+03 9077.374 8929.285 72.65799 72.65799 45.04457  
## ACF1 Theil's U  
## Training set 0.9945615 NA  
## Test set 0.9698456 45.12451

## [1] "Naive Method(naive)"

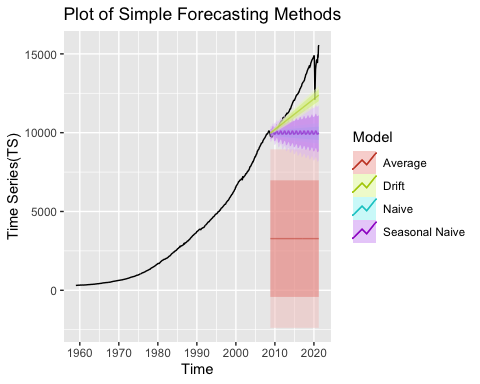
## ME RMSE MAE MPE MAPE MASE  
## Training set 16.17119 27.02136 18.29408 0.5802001 0.6543038 0.09228611  
## Test set 2247.72067 2778.27788 2272.06333 16.9391327 17.1882550 11.46162549  
## ACF1 Theil's U  
## Training set 0.1289840 NA  
## Test set 0.9698456 12.6108

## [1] "Seasonal Naive Method(snaive)"

## ME RMSE MAE MPE MAPE MASE  
## Training set 198.2322 246.7619 198.2322 6.798238 6.798238 1.00000  
## Test set 2212.9927 2751.9802 2243.0233 16.649098 16.955855 11.31513  
## ACF1 Theil's U  
## Training set 0.9831539 NA  
## Test set 0.9702406 12.47801

## [1] "Drift Method(rwf)"

## ME RMSE MAE MPE MAPE MASE  
## Training set -4.186415e-14 21.64824 15.16891 -0.8541177 1.142120 0.07652091  
## Test set 1.026796e+03 1401.77554 1091.42335 7.5258537 8.176922 5.50578213  
## ACF1 Theil's U  
## Training set 0.1289840 NA  
## Test set 0.9522344 6.2722



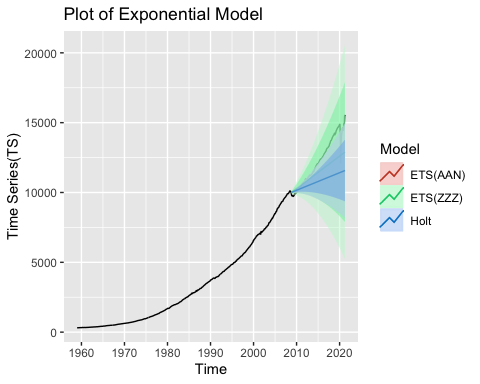
The analysis demonstrates that the drift method outperforms other models since the drift forecast value is closer to the original time series plot. When comparing MASE, RMSE, and MAE with other models, the performance metrics suggest that drift has the lowest errors. Drift’s MASE is 5.5057, which is the lowest of all other models.

The seasonal naive performance will not be as good as drift due to seasonaly adjusted data, whereas the naive and average methods will have a large error difference from the original time series data due to the same observation, which is the last observation and average value of periods for the forecast of future periods, respectively.

### 2

### Exponential Smoothing Model

Exponential smoothing is a univariate data time series forecasting approach that can be expanded to support data with a systematic trend or seasonal component (Brownlee, 2018). The method forecasts by analysing historical data, however recent data is more weighted in the times series dataset than older observations.



## [1] "ETS(Auto)"

## ME RMSE MAE MPE MAPE MASE  
## Training set 0.5223753 18.44674 10.57702 0.04864174 0.3851325 0.0533567  
## Test set 758.9523741 1119.39232 854.71972 5.44599427 6.3953767 4.3117096  
## ACF1 Theil's U  
## Training set -0.05219364 NA  
## Test set 0.94318255 4.985738

## [1] "Holt"

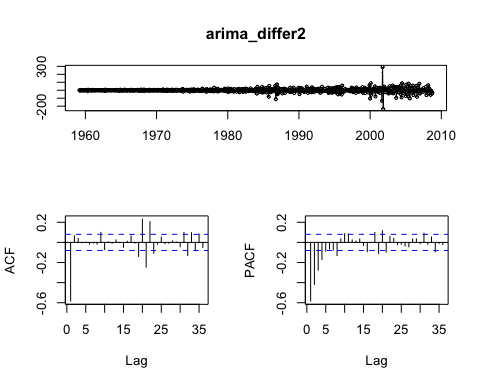
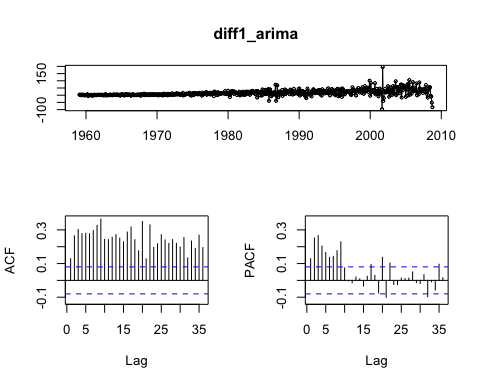
## ME RMSE MAE MPE MAPE  
## Training set 0.1810261 18.35107 10.55761 0.03285856 0.3877926  
## Test set 1414.9837017 1848.78416 1469.51195 10.49312172 11.0466890  
## MASE ACF1 Theil's U  
## Training set 0.05325878 0.02543583 NA  
## Test set 7.41308369 0.96169627 8.317797

Here is an illustration of an exponential model with AAN, where each letter stands for (error type, Trend error, Seasonal Error). As there is no seasonality in the time series, none of the errors are passed as seasonal errors, whereas error and trend type are additive, and the holt method is used for exponential smoothing when the time series is a trend type model. According to the graph, both methods are equivalent in terms of error performance measures, but the automatic model of seasonal smoothing surpasses other models with the least error.

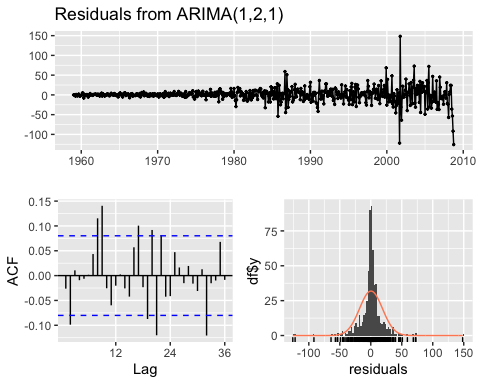
### 3

### ARIMA Model

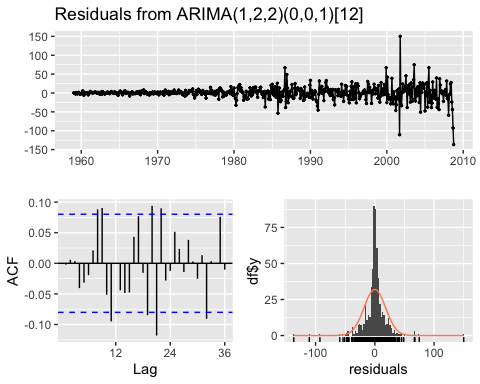
ARIMA stands for AutoRegressive Integrated Moving Average, and it uses the ARIMA function with the order (p,d,q), where d is the degree of differencing used to make the timeseries stationary, p is the number of lagged observations obtained from PACF, and q is the moving average order obtained from the ACF graph.



##   
## Augmented Dickey-Fuller Test  
##   
## data: arima\_differ2  
## Dickey-Fuller = -15.07, Lag order = 8, p-value = 0.01  
## alternative hypothesis: stationary



##   
## Ljung-Box test  
##   
## data: Residuals from ARIMA(1,2,1)  
## Q\* = 65.896, df = 22, p-value = 2.879e-06  
##   
## Model df: 2. Total lags used: 24



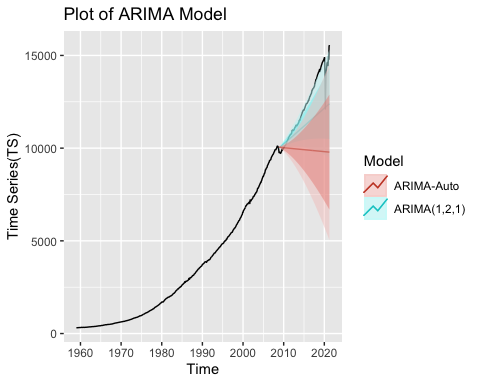
##   
## Ljung-Box test  
##   
## data: Residuals from ARIMA(1,2,2)(0,0,1)[12]  
## Q\* = 52.104, df = 20, p-value = 0.0001099  
##   
## Model df: 4. Total lags used: 24

## [1] "ARIMA(1,2,1)"

## ME RMSE MAE MPE MAPE MASE  
## Training set 0.3452932 18.4798 10.66186 0.03628569 0.3871204 0.05378471  
## Test set 1012.4359883 1393.0739 1083.41501 7.40308560 8.1175775 5.46538334  
## ACF1 Theil's U  
## Training set -0.02627709 NA  
## Test set 0.95234376 6.230874

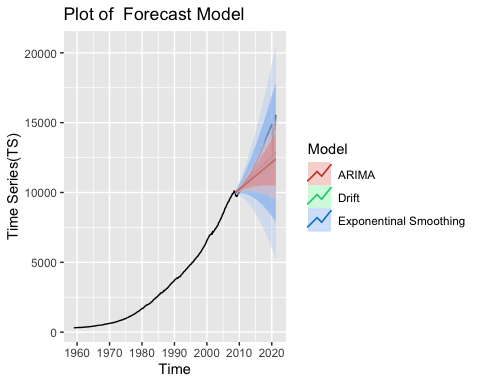
## [1] "AUTO-ARIMA"

## ME RMSE MAE MPE MAPE  
## Training set -0.05898444 18.13692 10.49465 0.01737654 0.3878492  
## Test set 2291.26466752 2857.90083 2329.82464 17.21833818 17.6122186  
## MASE ACF1 Theil's U  
## Training set 0.05294122 -0.002155045 NA  
## Test set 11.75300752 0.970820429 12.95921



Following the analysis, a model with a degree of differencing of two was created and tested for stationarity using adf.test, which confirmed stationarity. When compared to the auto-arima model, the ARIMA model passing order(1,2,1) has better performance metrics.

### Model Comparison



Exponential smoothing is considerably closer to the original timeseries plot and has the smallest error difference, as shown in the graph.

### Estimation of PCE (Oct 2021)

Estimating the PCE for the month october 2021 with the best forecast model i.e, exponential smoothing.

## [1] "Estimation of PCE(Oct 2021):15873.7183"

### One-Step Rolling Variation without Re-Estimation

This model is used to get test data by fitting the training and original timeseries dataset forecasts in a single step, a process known as single step rolling estimate.

## [1] "Drift Model"

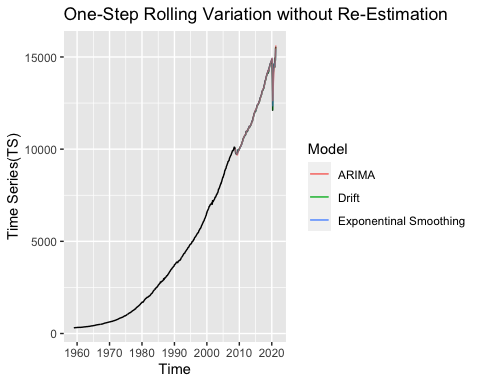
## ME RMSE MAE MPE MAPE ACF1 Theil's U  
## Test set 16.91272 216.632 74.07516 0.1132693 0.5762795 0.126176 0.9843011

## [1] "Exponential Model"

## ME RMSE MAE MPE MAPE ACF1 Theil's U  
## Test set 7.445166 229.3939 74.57784 0.03981773 0.5802979 0.2472555 1.035127

## [1] "Arima Model"

## ME RMSE MAE MPE MAPE ACF1 Theil's U  
## Test set 10.03898 236.685 81.50063 0.05832193 0.634292 0.3605788 1.059836

 When looking at the graph, all of the models perform about identically, with the exception of drift, which has somewhat better performance.

## Conclusion

The best possible future trend of personal consumption expenditure is discovered using exponential smoothing, which is visualised using forecasting model analysis. The data that is varying over time is discovered by the output of future time series, and the forecast value increases gradually over time. One of the major drawbacks is the unpredictability of forecast data, which makes it impossible to provide a precise output because a lot of things might change in a day’s time. As a result, the forecast provided is a preliminary study of probable solutions.