MIDTERM
2016FA_PREDICT_413-DL_SEC59
NITIN GAONKAR

Weather forecasting has been one of the most scientifically and technologically challenging problems around the world in the last century. This is due mainly to two factors: first, it's used for many human activities and secondly, due to the opportunism created by the various technological advances that are directly related to this concrete research field, like the evolution of computation and the improvement in measurement systems. To make an accurate prediction is one of the major challenges facing meteorologist all over the world. Since ancient times, weather prediction has been one of the most interesting and fascinating domain. Scientists have tried to forecast meteorological characteristics using a number of methods, some of these methods being more accurate than others. Since the oldest human civilization, humans have attempted to predict the weather informally. Now, weather forecasting is made through the application of science and technology. It is made by collecting quantitative data about the current state of the atmosphere through weather station and interprets by meteorologist **[1].** Multivariate techniques have been underlined as

suitable and powerful tools for classifying the meteorological data such as rainfall. Principal components, factor analysis and different cluster techniques have been used to classify daily rainfall patterns and their relationship to the atmospheric circulation.

On the other hand, time series modeling is a major tool in planning, operating and decision making of water resources and investigating climatic fluctuations and has been commonly used for data generation, forecasting, estimating missing data and extending hydrologic data records. To accomplish these objectives, the modeler has to decide on choosing the type of the model whether it is univariate or multivariate. These model types are generally based on the annual time series with homogenous mean and variance or on the seasonal series generally with periodic parameters [2].

This paper addresses precipitation forecasts by using various techniques like regression, arima models, ets,naive etc., we had access to around 66 years of daily meteorological data and to forecast rainfall for 24 months based on the historical meteorological data.

Weather forecasts are based on complex models where the variables are determined through observations such as precipitation amount and intensity. So, precipitation measurement is important for weather forecasting; accurate forecasting is most important from agriculture and water resource perspective, an accurate forecast week in advance can provide a distinct reduction in agricultural vulnerability to climatic variations and ability to take advantage of favorable future conditions, also the experts feels that the accurate forecast could increase the agriculture yield by 20 to 30% [1]

The data (meteorological daily data) for this project was obtained for about 66 years from the meteorological department, the data was provided from the GHCN daily, it's a database that addresses the critical need historical daily temperature, precipitation snow records over global land areas. GHCN-Daily is a composite of climate records from numerous sources that were merged and then subjected to a suite of quality assurance reviews. The archive includes over 40 meteorological elements including temperature daily maximum/minimum, temperature at observation time, precipitation, snowfall, snow depth, evaporation, wind movement, wind maximums, soil temperature, cloudiness. Due to the time constraints only the daily precipitation is provided. The precipitation is the daily precipitation from 1946 September to July 2014. Before starting with the modeling below is the summary of the data what I used for the forecasting purpose, the daily precipitation data was provided and I aggregated to monthly data for all the analysis, below is the summary of the original data and the aggregated data.

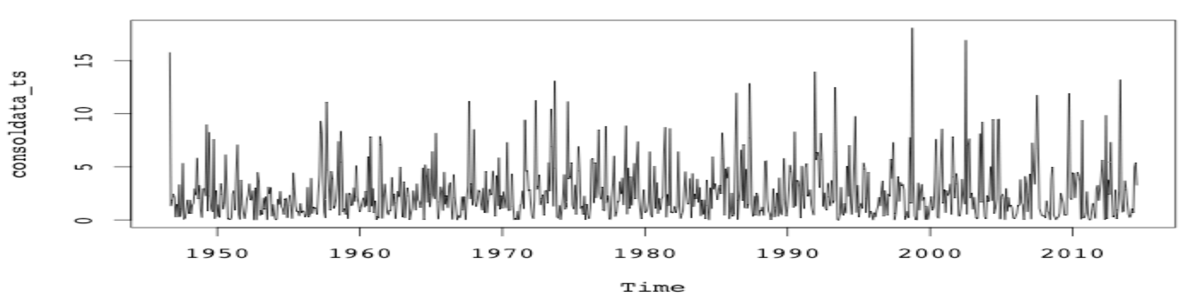
Original data:

```
> summary(original_data)
      DATE                  PRCP
 Min.    :19460901   Min.    : 0.00000000
 1st Qu.:19630824   1st Qu.: 0.00000000
 Median :19800816   Median : 0.00000000
 Mean    :19801901   Mean    : 0.08342659
 3rd Qu.:19970808   3rd Qu.: 0.00000000
 Max.    :20140731   Max.    :11.26000000
```

Aggregated to monthly data:

```
     Index              consoldata_bkp
 Min.    :1946-09-30   Min.    : 0.000000
 1st Qu.:1963-09-15   1st Qu.: 0.735000
 Median :1980-08-31   Median : 1.860000
 Mean    :1980-08-30   Mean    : 2.539239
 3rd Qu.:1997-08-15   3rd Qu.: 3.380000
 Max.    :2014-07-31   Max.    :18.070000
```

Below is the plot for the timseries data:



1. Types of model:

I started off with a simple model using the average method naïve and the drift method.

Average/naïve/drift method:

The data set was aggregated on monthly basis and below is the sample data after the aggregation and converted into a time series.

```
> head(consoldata_ts)
1946-09-30 1946-10-31 1946-11-30 1946-12-31 1947-01-31 1947-02-28
     15.78       1.31       1.86       2.43       2.14       0.29
```

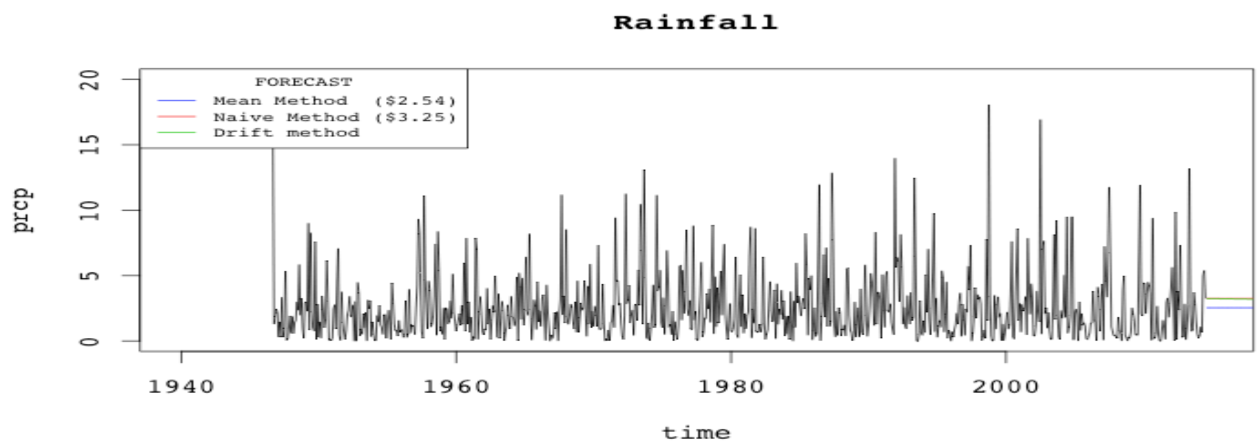
Below is the forecast for the 24 months based on the average method.

```
> meanf(consoldata_ts)
         Point Forecast         Lo 80       Hi 80       Lo 95       Hi 95
Aug 2014    2.539239264 -0.7799005098 5.858379037 -2.54038024 7.618858768
Sep 2014    2.539239264 -0.7799005098 5.858379037 -2.54038024 7.618858768
Oct 2014    2.539239264 -0.7799005098 5.858379037 -2.54038024 7.618858768
Nov 2014    2.539239264 -0.7799005098 5.858379037 -2.54038024 7.618858768
Dec 2014    2.539239264 -0.7799005098 5.858379037 -2.54038024 7.618858768
Jan 2015    2.539239264 -0.7799005098 5.858379037 -2.54038024 7.618858768
Feb 2015    2.539239264 -0.7799005098 5.858379037 -2.54038024 7.618858768
Mar 2015    2.539239264 -0.7799005098 5.858379037 -2.54038024 7.618858768
Apr 2015    2.539239264 -0.7799005098 5.858379037 -2.54038024 7.618858768
May 2015    2.539239264 -0.7799005098 5.858379037 -2.54038024 7.618858768
```

Now tried forecasting with the naïve method:

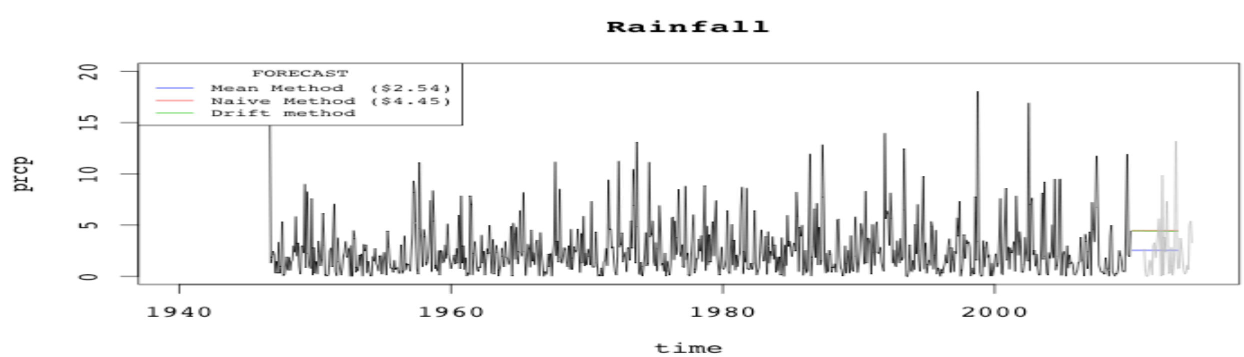
```
> naive(consoldata_ts)
         Point Forecast        Lo 80       Hi 80        Lo 95       Hi 95
Aug 2014           3.25 -1.167518563 7.667518563 -3.506011632 10.00601163
Sep 2014           3.25 -1.167518563 7.667518563 -3.506011632 10.00601163
Oct 2014           3.25 -1.167518563 7.667518563 -3.506011632 10.00601163
Nov 2014           3.25 -1.167518563 7.667518563 -3.506011632 10.00601163
Dec 2014           3.25 -1.167518563 7.667518563 -3.506011632 10.00601163
Jan 2015           3.25 -1.167518563 7.667518563 -3.506011632 10.00601163
Feb 2015           3.25 -1.167518563 7.667518563 -3.506011632 10.00601163
Mar 2015           3.25 -1.167518563 7.667518563 -3.506011632 10.00601163
Apr 2015           3.25 -1.167518563 7.667518563 -3.506011632 10.00601163
May 2015           3.25 -1.167518563 7.667518563 -3.506011632 10.00601163
```

Below is the plot for the mean/Naïve and drift method.

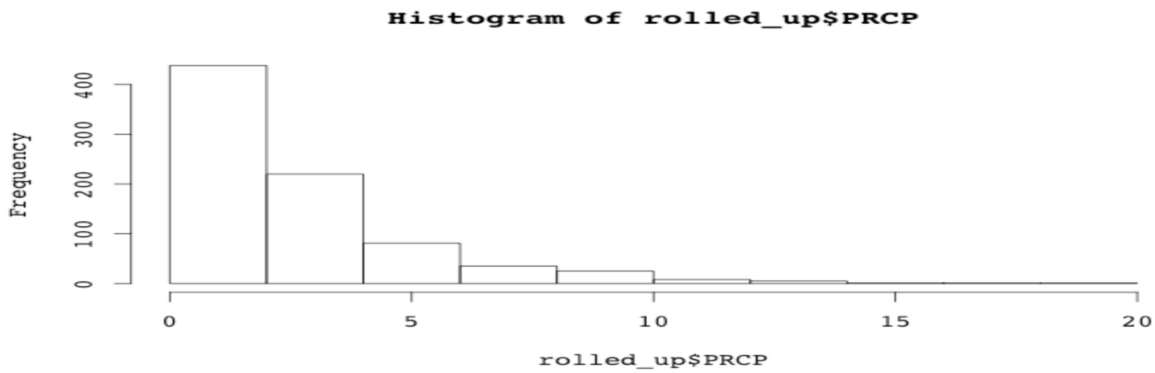


The data was divided into sets of test and train and was later plotted to see if we can forecast and compare with the actual data. Divided the train data until 2010 and made the remaining data as test, below are the results.

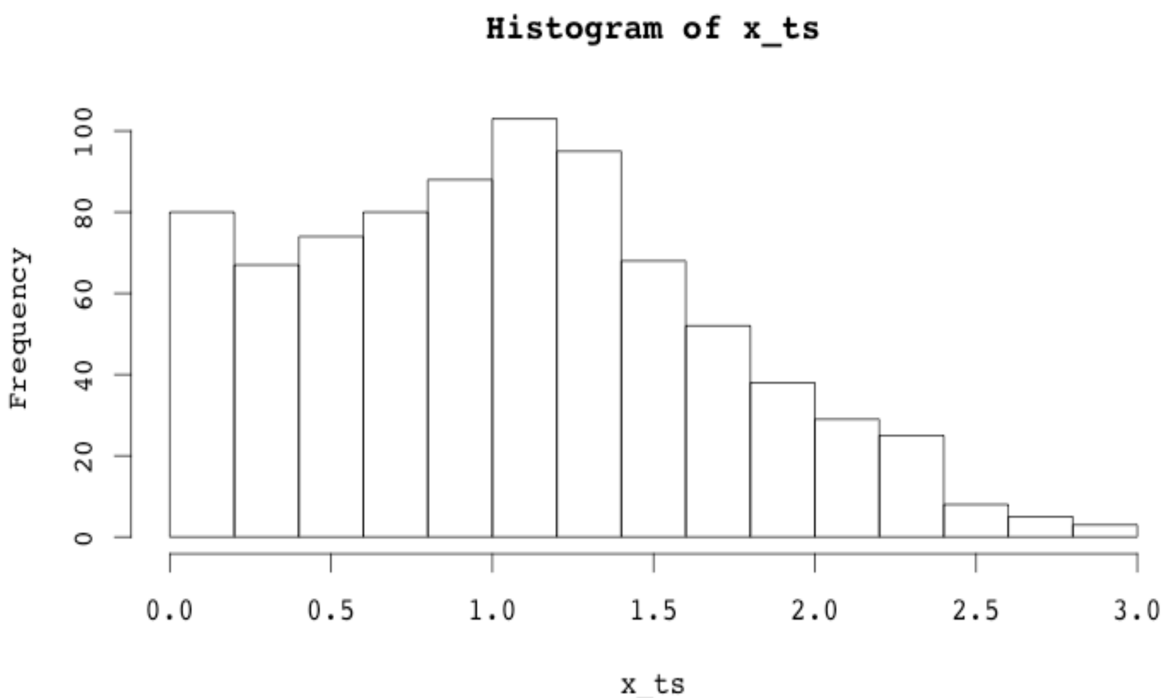
The naïve and draft method looks somewhat better than the average method but nothing conclusive yet.



Before proceeding with the any further modelling, I tried to plot the PRCP on the histogram to see if the data is normal:



Added +1 to the time series and log transformed the PRCP



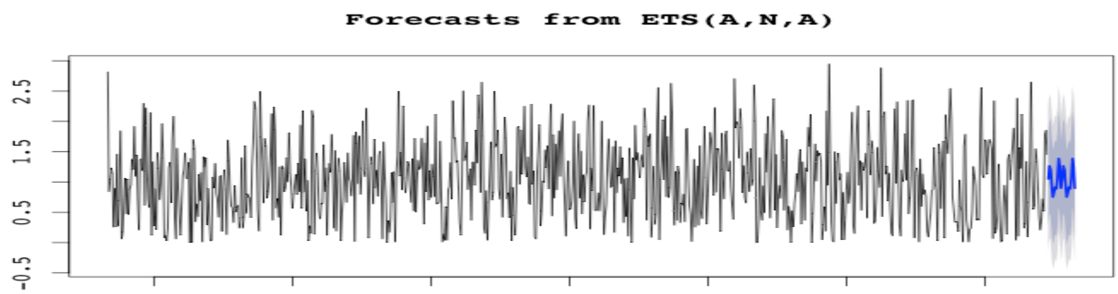
1.1 ETS/ARIMA/TBATS METHODS:

ETS:
Below is the plot from the ETS method and the forecast:

All the related code is in the R file.

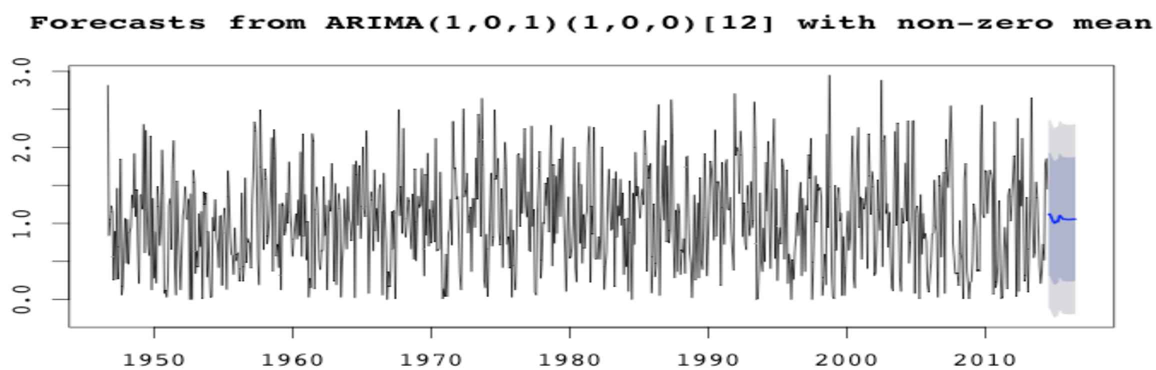
		Point Forecast	Lo 80	Hi 80	Lo95	Hi 95
Aug	2014	2.855567444	1.319856156	6.17814706	0.877203862	9.2957473
Sep	2014	3.55083098	1.641171666	7.682560523	1.090743314	11.5594572
Oct	2014	3.323784292	1.536196146	7.191491816	1.020962608	10.82071168
Nov	2014	2.590262428	1.197146877	5.604541577	0.795619133	8.433004144
Dec	2014	2.11630264	0.978073011	4.579143697	0.650015442	6.890200722
Jan	2015	2.299968048	1.062931152	4.976665717	0.706402375	7.488441729
Feb	2015	2.501477552	1.156031781	5.412818268	0.768265633	8.144826049
Mar	2015	2.422016599	1.119283524	5.240999519	0.743834573	7.8863831
Apr	2015	2.88630159	1.333811742	6.245811618	0.886391134	9.398488496
May	2015	4.013672845	1.854746798	8.685589708	1.232565804	13.06994698
Jun	2015	3.450980378	1.594684304	7.46810233	1.059728667	11.23803284
Jul	2015	2.446824263	1.130641289	5.295179853	0.751345037	7.968308405
Aug	2015	2.855567444	1.319484576	6.179886886	0.876826197	9.299751143
Sep	2015	3.55083098	1.64070964	7.684723944	1.090273728	11.56443591
Oct	2015	3.323784292	1.535763686	7.193516894	1.020523077	10.82537207
Nov	2015	2.590262428	1.196809874	5.60611973	0.795276625	8.436636051
Dec	2015	2.11630264	0.977797687	4.580433073	0.649735623	6.893168095
Jan	2016	2.299968048	1.062631949	4.978066985	0.706098292	7.491666647
Feb	2016	2.501477552	1.155706382	5.414342294	0.76793493	8.148333534
Mar	2016	2.422016599	1.118968478	5.24247512	0.743514397	7.889779179
Apr	2016	2.88630159	1.333436324	6.247570074	0.886009606	9.402535604
May	2016	4.013672845	1.854224772	8.688034994	1.23203529	13.07557489
Jun	2016	3.450980378	1.594235488	7.470204784	1.059272559	11.24287179
Jul	2016	2.446824263	1.130323085	5.296670529	0.751021667	7.971739349

The plot looks ok; the forecast is shown in blue with the grey area representing a 95% confidence interval. Just by looking, we see that the forecast roughly matches the historical pattern of the data.



1.2 ARIMA:

Below are the plots and forecast:

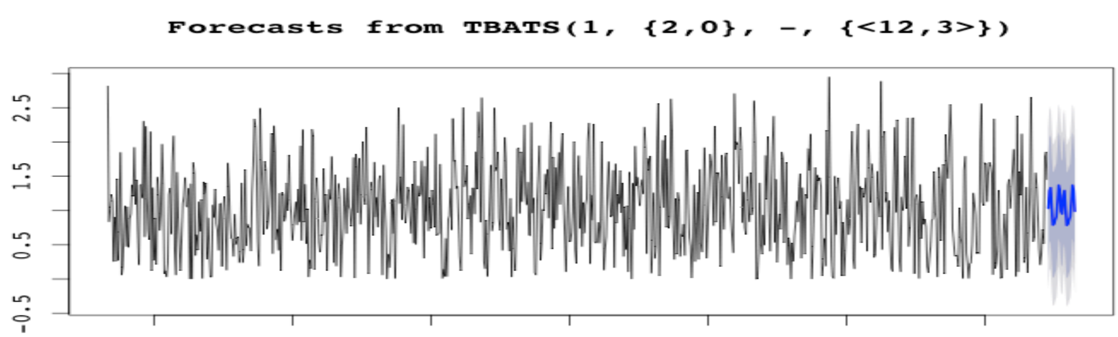


		Point forecast	lo 80	Hi 80	lo 95	Hi 95
Aug	2014	3.053277223	1.363973935	6.834809346	0.890324551	10.47090276
Sep	2014	3.072277791	1.363865766	6.92068902	0.887297846	10.63779302
Oct	2014	2.967635575	1.315829237	6.693012027	0.855501753	10.2943809
Nov	2014	2.865323387	1.270171327	6.463756453	0.82571577	9.942983296
Dec	2014	2.772036878	1.22876373	6.253593159	0.798778656	9.61992211
Jan	2015	2.728529081	1.20946766	6.155494016	0.786231367	9.46905868
Feb	2015	2.750071282	1.219014613	6.204102874	0.792436809	9.54384246
Mar	2015	2.811570702	1.246274855	6.342846262	0.81015756	9.757274633
Apr	2015	2.776720119	1.230826687	6.264224439	0.800115247	9.63633007
May	2015	2.997531342	1.328704864	6.762370174	0.863742251	10.40263357
Jun	2015	3.009491197	1.334006263	6.789351385	0.867188495	10.4441391
Jul	2015	2.936503362	1.301653209	6.624692303	0.846156962	10.19084209
Aug	2015	2.878411458	1.274047519	6.503095374	0.827573707	10.01149803
Sep	2015	2.879482968	1.274492747	6.505664461	0.827852923	10.01557395

Oct	2015	2.873458869	1.271820838	6.492082555	0.826115455	9.994687572
Nov	2015	2.867374578	1.269126803	6.478341609	0.824365168	9.973537565
Dec	2015	2.861647122	1.266591572	6.465402451	0.822718331	9.953618327
Jan	2016	2.858913727	1.265381707	6.459227012	0.821932445	9.944111282
Feb	2016	2.860271954	1.265982862	6.462295731	0.822322925	9.948835658
Mar	2016	2.864095441	1.267675172	6.470934254	0.82342217	9.96213485
Apr	2016	2.861938453	1.266720469	6.466060915	0.822802039	9.954632244
May	2016	2.875196627	1.27258866	6.496015494	0.82661374	10.00074793
Jun	2016	2.875888247	1.272894777	6.497578084	0.826812579	10.00315358
Jul	2016	2.871626572	1.27100852	6.487949557	0.825587355	9.988330255

1.3 TBATS:

Plot for TBATS:

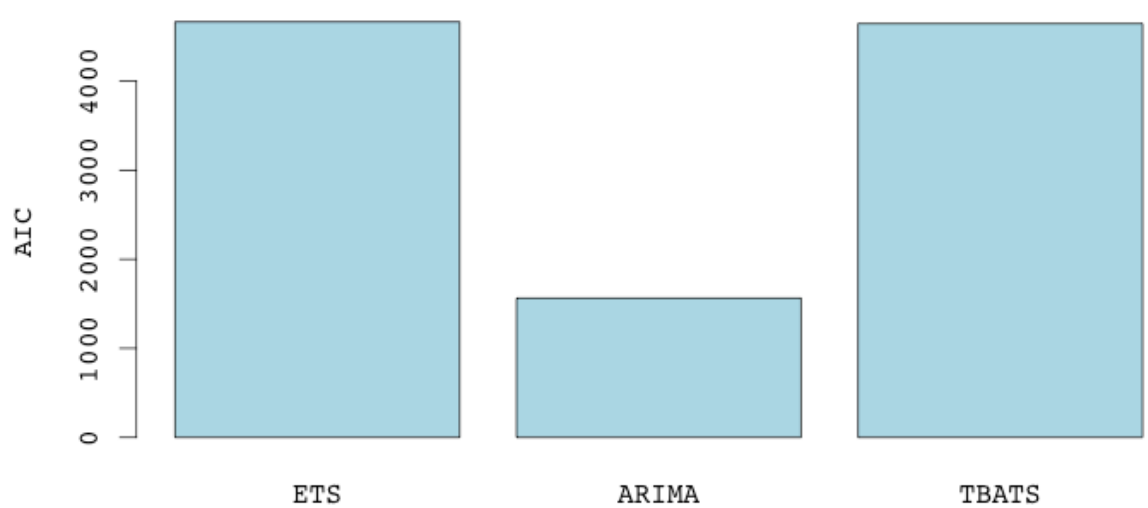


		Point forecast	lo 80	Hi 80	lo 95	Hi 95
Aug	2014	2.805224462	1.307294064	6.019521156	0.872645504	9.017733143
Sep	2014	3.626863443	1.685922936	7.802336725	1.123881252	11.70420666
Oct	2014	3.793519453	1.757540571	8.188027104	1.169563858	12.30440711
Nov	2014	2.790167517	1.292344281	6.023963492	0.859876098	9.053670395
Dec	2014	2.195311684	1.01672251	4.74012658	0.676453607	7.124499507
Jan	2015	2.229885095	1.03270832	4.814900238	0.687080147	7.236983276
Feb	2015	2.373539841	1.099199731	5.125266338	0.731304662	7.703617479
Mar	2015	2.483989958	1.150321701	5.363896125	0.765306541	8.062398251
Apr	2015	3.06684414	1.420226166	6.622559986	0.944869168	9.954323098
May	2015	3.938354888	1.823790581	8.504616366	1.213350092	12.7833173
Jun	2015	3.644761332	1.687796389	7.870786579	1.122861942	11.83073775

Jul	2015	2.684098566	1.242936464	5.796261774	0.826903783	8.712482958
Aug	2015	2.583215292	1.19620904	5.578457464	0.795812976	8.385137521
Sep	2015	3.371860164	1.561406451	7.281538359	1.038771133	10.94508752
Oct	2015	3.640247414	1.685621943	7.861431378	1.12138568	11.8169881
Nov	2015	2.765349302	1.280412455	5.972416725	0.85178339	8.977818607
Dec	2015	2.186818545	1.012513392	4.723073675	0.673556097	7.099891711
Jan	2016	2.227659522	1.03140609	4.811360907	0.686118141	7.23267124
Feb	2016	2.372644435	1.098501255	5.124656513	0.730740027	7.703754289
Mar	2016	2.483724118	1.149902755	5.364701896	0.764923647	8.064707523
Apr	2016	3.066729036	1.419807934	6.624013546	0.944462422	9.957862548
May	2016	3.938310648	1.823301621	8.506705961	1.212859835	12.78819722
Jun	2016	3.644747549	1.687356538	7.872778747	1.122416688	11.8353414
Jul	2016	2.684095436	1.2426158	5.797744009	0.826578052	8.715895966

Now comparing these three models for by the AIC

Below plot compares the AIC values of the three models which we built, by looking at the values it looks like the ARIMA model is better of the other two.



After comparing the other factors like RMSE, MAD, MSE for the above models along with the AIC and other important factors, I am picking ETS model for submission.

Below are the rmse for the models

```
> c=c(rmse_ets,ma_rmse)
> c
[1] 0.6021980236 0.6272343596
>
```

1.4 Regression model:

In this regression model, I have transformed the time series with log transformation and ran a regression with trend and season.

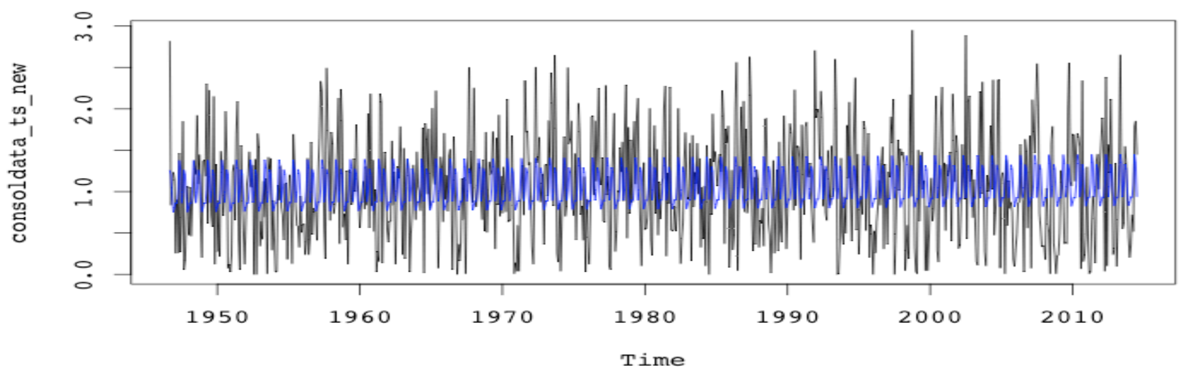
```
> fit<-tslm(consoldata_ts_new ~ trend + season)
> summary(fit)
```

Below is the summary of the fit.

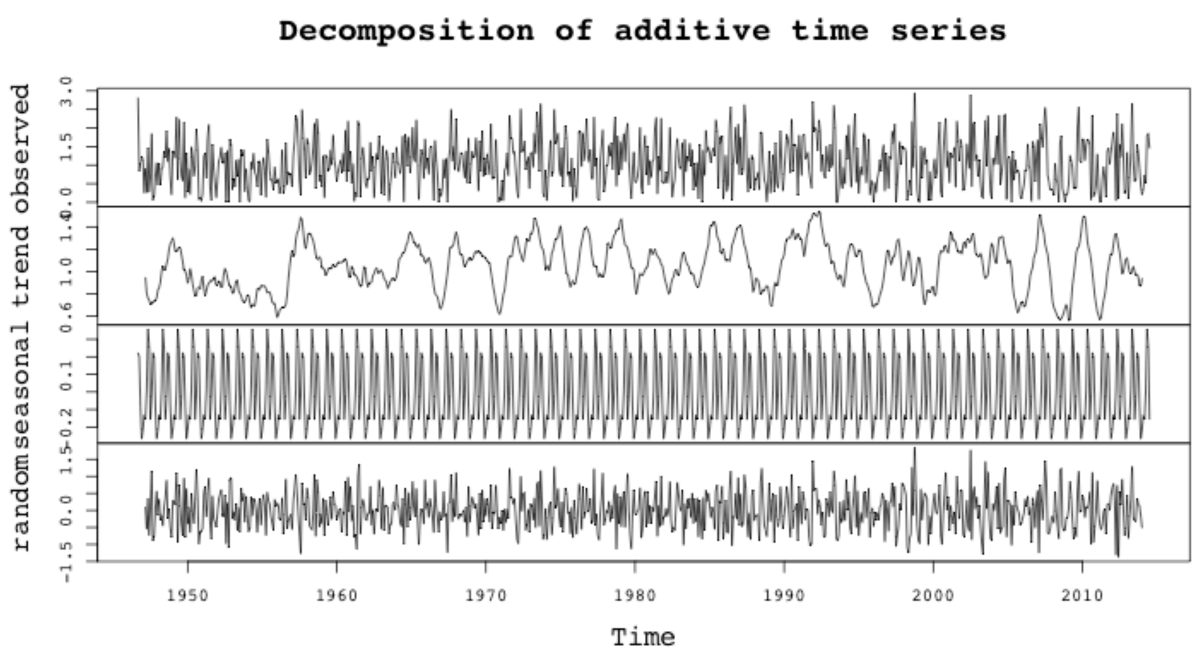
```
Call:
tslm(formula = consoldata_ts_new ~ trend + season)

Residuals:
        Min          1Q      Median          3Q         Max
-1.33384956 -0.44592380 -0.00913048  0.41056292  1.96232132

Coefficients:
                Estimate     Std. Error  t value                Pr(>|t|)
(Intercept)  0.80488849015 0.08212399156  9.80089 < 0.0000000000000000222 ***
trend        0.00009227207 0.00009019992  1.02297              0.30662912
season2      0.05924784087 0.10389061183  0.57029              0.56864033
season3      0.04974741950 0.10389072930  0.47884              0.63218036
season4      0.21031260210 0.10389092508  2.02436              0.04326414 *
season5      0.57096427448 0.10389119917  5.49579          0.000000052278 ***
season6      0.47044552286 0.10389155158  4.52824          0.000006850301 ***
season7      0.05679304039 0.10389198230  0.54665              0.58476803
season8      0.18123116198 0.10427754253  1.73797              0.08260006 .
season9      0.45380036205 0.10389119917  4.36803          0.000014181354 ***
season10     0.39850331057 0.10389092508  3.83579              0.00013499 ***
season11     0.12254420843 0.10389072930  1.17955              0.23852940
season12    -0.05200620285 0.10389061183 -0.50059              0.61679974
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6057809 on 802 degrees of freedom
Multiple R-squared:  0.1020502,	Adjusted R-squared:  0.08861452
F-statistic: 7.595473 on 12 and 802 DF,  p-value: 0.000000000000001995005
```

Apart from the coefficient estimates and their standard error, the output also includes the corresponding t-statistics and p–values. In our case, the coefficients intercept and the season 5, 6, 9 and 10 looks are significant rest doesn't look like having any impact on the model.

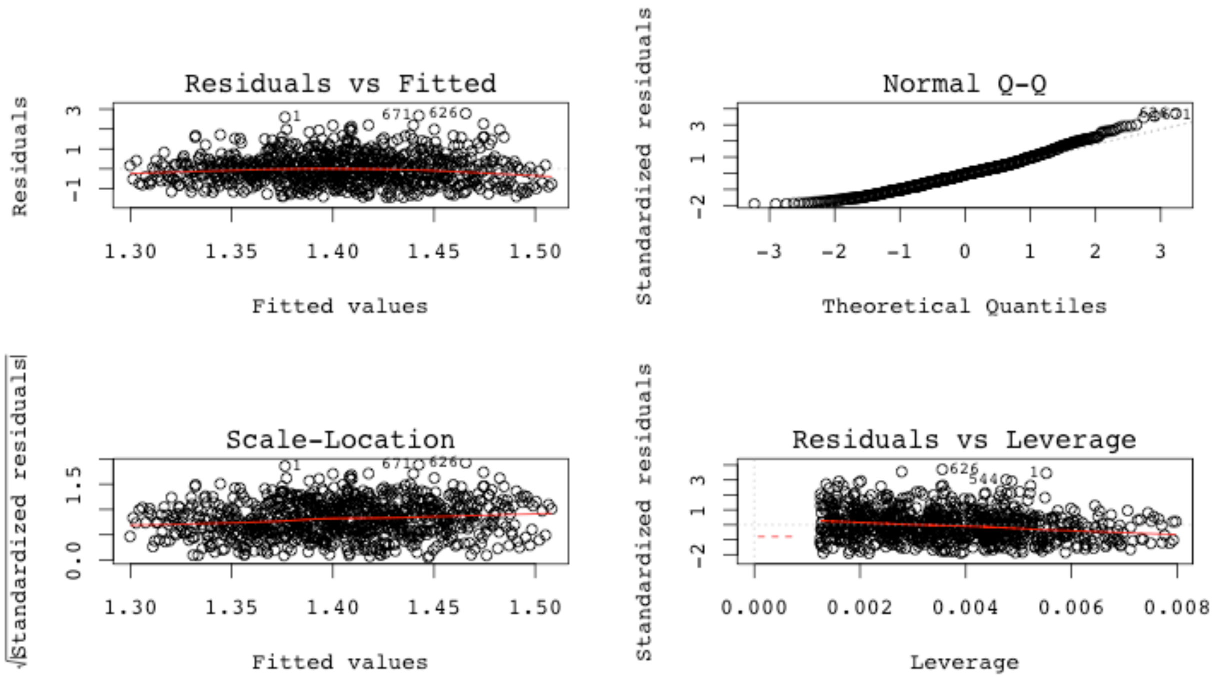
Below is the plot of time series



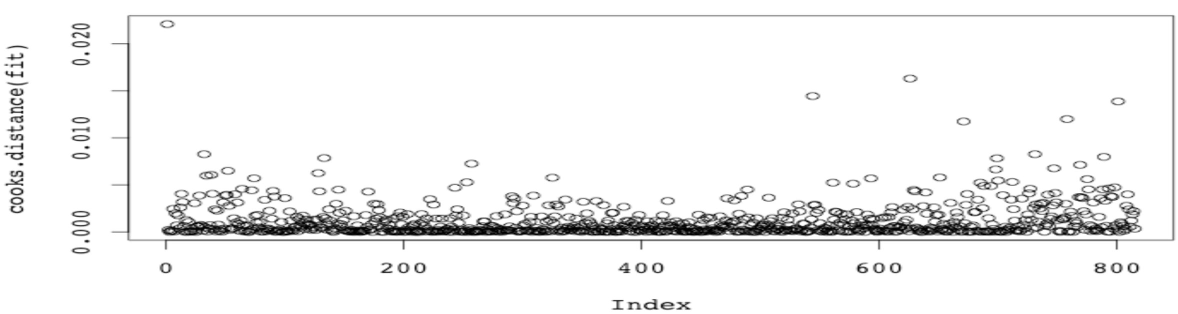
Based on the decompose chart, we do see a trend and strong seasonality.

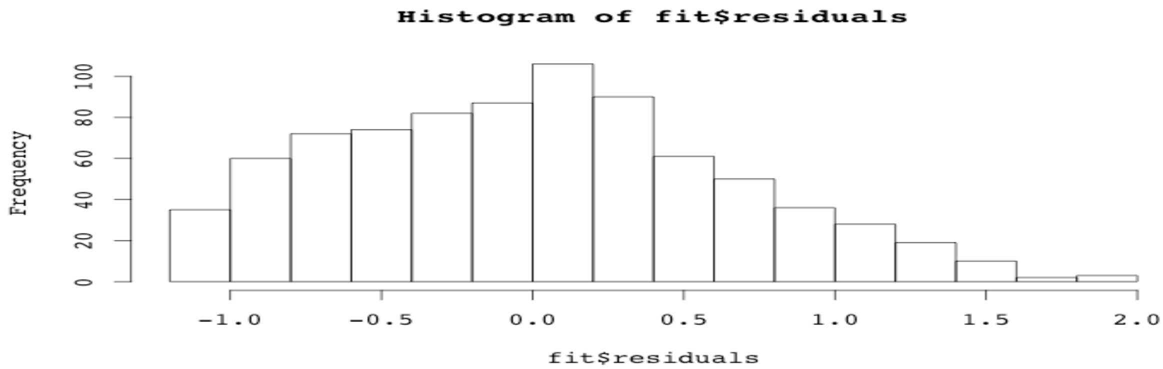


Residual plots:



from the above residual plot we can see that the Normal qq plot looks ok,the residual vs leverage plot looks good as well. below is the cook’s plot, we can see that there are few outliers in the graph.





Above is the residual histogram, looks like its slightly positively skewed may be because of the outliers.

Below is the sample test data for prediction

```
> head(new_data)
  Year Month
1 2014     8
2 2014     9
3 2014    10
4 2014    11
5 2014    12
6 2015     1
```

Now ran the forecast on the model.

frct <- forecast.lm(fit, newdata=new_data)

		Point forecast	lo 80	Hi 80	lo 95	Hi 95
Aug	2014	2.890454231	1.319476921	6.331846761	0.870483203	9.597802269
Sep	2014	3.796483089	1.733147737	8.316246523	1.143415166	12.60546849
Oct	2014	3.592579097	1.640062707	7.869592129	1.082003878	11.92844576
Nov	2014	2.726458532	1.24466653	5.972343551	0.821147879	9.052664353
Dec	2014	2.289985273	1.045410371	5.016243095	0.689691968	7.603441532
Jan	2015	2.412452492	1.101318372	5.284509161	0.726576334	8.010069623
Feb	2015	2.559940394	1.168648667	5.607583365	0.770996367	8.499773924
Mar	2015	2.535969032	1.157705404	5.555073776	0.763776733	8.420181779
Apr	2015	2.977944448	1.359473376	6.523226784	0.89688977	9.887673417
May	2015	4.271551988	1.950023347	9.356891241	1.286495214	14.18284047
Jun	2015	3.863412054	1.763701747	8.462855312	1.163572662	12.82769283

Jul	2015	2.554842389	1.166321356	5.596416118	0.769460963	8.482847012
Aug	2015	2.893656502	1.320826743	6.339399165	0.871334518	9.609682376
Sep	2015	3.80068913	1.734918604	8.326176132	1.144531242	12.62109529
Oct	2015	3.596559237	1.641738463	7.878988424	1.08306001	11.94323326
Nov	2015	2.729479115	1.245938285	5.979474531	0.821949392	9.063886794
Dec	2015	2.292522298	1.046478533	5.02223249	0.690365169	7.612867403
Jan	2016	2.415125195	1.102443659	5.290818866	0.727285538	8.019999584
Feb	2016	2.562776496	1.169842749	5.614278821	0.771748929	8.510310956
Mar	2016	2.538778577	1.158888305	5.561706535	0.764522248	8.430620142
Apr	2016	2.981243647	1.360862436	6.531015518	0.897765215	9.899931006
May	2016	4.276284347	1.952015809	9.368063381	1.287750949	14.2004227
Jun	2016	3.867692244	1.765503833	8.472959963	1.164708413	12.84359512
Jul	2016	2.557672843	1.16751306	5.603098241	0.770212026	8.49336306

Along with these models I also tried few other regression models which are not

documented in this paper. Tried few of the models with different variables but the one with trend and season was better than all other regression models after comparing the MAD, MSE and RMSE

t<-seq(1946.7,2014.7.2,length=length(consoldata_ts))

t2<-t^2
sin.t<-sin(2*pi*t)
cos.t<-cos(2*pi*t)
plot(consoldata_ts)

Models tried were:

> fit<-tslm(consoldata_ts_new ~ t)
> fit<-tslm(consoldata_ts_new ~ t + t2)
> fit<-lm(consoldata ~ Month + Year)

2. Implementation:

I have created functions to implement the models and I will be choosing Regression

model and the ETS model for the final submission, but the functions are created for 4 models

First function for Regression:

```
#forecasting regression
forecastit_reg=function(model,forecast,level)
    {
    fit<-tslm(model)     #runregression
    frct <- forecast.lm(fit,forecast)
    mylist=list(frct)
    return(mylist)
    }
#example
consoldata_ts
consoldata_ts_new<- consoldata_ts + 1
consoldata_ts_new >- log(consoldata_ts_new)
model=consoldata_ts_new ~ trend + season
forecast=data.frame(new_data)
l=.05
forecastit_reg(model, forecast, level)
```

Naïve method:

```
#forecasting naive method
forecastit_naive=function(model,forecast,level)
    {
    fit<-naive(model)   #naive method
    forecast_df = data.frame(forecast_predicted=fit$mean,forecast_lower=fit$lower[,2],forecast_upper=fit$upper[,2])  # high 95%
    forecast_df_2 = data.frame(forecast_predicted=fit$mean,forecast_lower=fit$lower[,1],forecast_upper=fit$upper[,1])  # high 80%
    mylist=list(forecast_df,forecast_df_2)
    return(mylist)
    }
#example
consoldata_ts
consoldata_ts_new<- consoldata_ts + 1
consoldata_ts_new >- log(consoldata_ts_new)
model=consoldata_ts_new
forecast=data.frame(new_data)
l=.05
forecastit_naive(model, forecast, level)
```

ARIMA model:

```
#forecasting Arima method:
forecastit_arima=function(model,forecast,level)
    {
    fit<-auto.arima(model)  #runregression
    forecast_df = data.frame(forecast_predicted=f_aa$mean,forecast_lower=f_aa$lower[,2],forecast_upper=f_aa$upper[,2])  # high 95%
    forecast_df_2 = data.frame(forecast_predicted=f_aa$mean,forecast_lower=f_aa$lower[,1],forecast_upper=f_aa$upper[,1])  # high 80%
    mylist=list(forecast_df,forecast_df_2)
    return(mylist)
    }
#example
consoldata_ts
consoldata_ts_new<- consoldata_ts + 1
consoldata_ts_new >- log(consoldata_ts_new)
model=consoldata_ts_new
forecast=data.frame(new_data)
l=.05
forecastit_arima(model, forecast, level)
```

ETS method:

```
#forecasting ETS method:
forecastit_ets=function(model,forecast,level)
    {
    fit<-ets(model) #ets
    forecast_df = data.frame(forecast_predicted=f_ets$mean,forecast_lower=f_ets$lower[,2],forecast_upper=f_ets$upper[,2])  # high 95%
    forecast_df_2 = data.frame(forecast_predicted=f_ets$mean,forecast_lower=f_ets$lower[,1],forecast_upper=f_ets$upper[,1])  # high 80%
    mylist=list(forecast_df,forecast_df_2)
    return(mylist)
    }
#example
consoldata_ts
consoldata_ts_new<- consoldata_ts + 1
consoldata_ts_new >- log(consoldata_ts_new)
model=consoldata_ts_new
forecast=data.frame(new_data)
l=.05
forecastit_ets(model, forecast, level)
```

3. PEER REVIEW:

Case 1: [1] I came across this paper where they have used multiple regression models for forecasting.

In this paper, the team uses data mining technique in forecasting monthly Rainfall of Assam.

This was carried out using traditional statistical technique -Multiple Linear Regression. The data include Six years' period [2007-2012] collected locally from Regional Meteorological Center, Guwahati, Assam, India. The performance of their model is measured in adjusted R-squared Their experiments results show that the prediction model based on Multiple linear regression indicates acceptable accuracy.

Below was the approach followed in the journal:
- Data Collection
- Reduction explanatory predictors
- Building model using backward procedure - Validity Check

Below were the predictor variables used in the experiment and the results of the MLR

Table 1: Details of Predictors

Attributes	Type	Description
Year	Numeric	Year Considered
Rainfall	Numeric	Monthly rainfall considered
Min Temperature	Numeric	Min temperature in degree celcius
Max Temperature	Numeric	Max temperature in degree celcius
Relative Humidity	Numeric	Relative humidity in %
Wind Speed	Numeric	Wind run in Kmph
Pressure	Numeric	Mean sea level pressure in mb
Month	Numeric	Month Considered

Table 2: MLR Results

Regression Statistics	
Multiple R	0.842330832
R Square	0.709521231
Adjusted R Square	0.628832684
Standard Error	56.69918995
Observations	24

	Df	SS	MS	F	Significance F
Regression	5	141343.9468	28268.78936	8.793332619	0.000230647
Residual	18	57866.36654	3214.798141		
Total	23	199210.3133			

In this MLR implementation and they found that the models explain 63 % of the prediction of the rainfall. Here the approach and the model is picked based on the Adjusted R square value and the only the significant variables were used based on the P value.

Case 2: [2]

ARIMA Method:

Raymond Y.C. Tse, (1997) suggested that the following two questions must be answered to identify the data series in a time series analysis: (1) whether the data are random; and (2) have any trends? This is followed by another three steps of model identification, parameter estimation and testing for model validity. If a series is random, the correlation between successive values in a time series is close to zero. If the observations of time series are statistically dependent on each another, then the ARIMA is appropriate for the time series analysis. Meyler et al (1998) drew a framework for ARIMA time series models for forecasting Irish inflation. In their research, they emphasized heavily on optimizing forecast performance while focusing more on minimizing out-of-sample forecast errors rather than maximizing in-sample 'goodness of fit'. Stergiou (1989) in his research used ARIMA model technique on a 17 years' time series data (from 1964 to 1980 and 204 observations) of monthly catches of pilchard (Sardina pilchardus) from Greek waters for forecasting up to 12 months ahead and forecasts were compared with actual data for 1981 which was not used in the estimation of the parameters. The research found mean error as 14% suggesting that ARIMA procedure was capable of forecasting the complex dynamics of the Greek pilchard fishery, which, otherwise, was difficult to predict because of the year-to-year changes in oceanographic and biological conditions. Contreras et al (2003) in their study, using ARIMA methodology, provided a method to predict next-day electricity prices both for spot markets and long-term contracts for mainland Spain and Californian markets. In fact, a plethora of research studies is available to justify that a careful and precise selection of ARIMA model can be fitted to the time series data of single variable (with any kind of pattern in the series and with autocorrelations between the successive values in the time series) to forecast, with better accuracy, the future values in the series.

CASE3: [3]

The Prediction of Indian Monsoon Rainfall: A Regression Approach

In this Journal, multiple linear regression is used to predict the average summer-monsoon rainfall using the previous years' data from the corresponding time period.

In this research paper, a multiple linear regression (MLR) method is adopted to predict the average summer monsoon rainfall in a given year using the monthly rainfall data of the summer-monsoon of the previous year. After computation, the MLR equation is set as

$y=0.03x1+0.06x2+0.02x3+229$

Where, x1= June rainfall of year Y

x2= July rainfall of year Y

x3= August rainfall of year Y

y= Average rainfall of year Y+1

This journal helped me pick the predictor variables based on the months and use it in my regression model and again Adjusted R square component and the residuals plots are used to select the model for prediction.

CASE4: [4]

Brandt, J.A., and Bessler, D.A. (1983), "Price forecasting and evaluation: An application in agriculture", Journal of Forecasting, 2, 237-248. F, TS, R, EO, A, E. The authors show the economic impact of various forecasting strategies by calculating average prices obtained for hogs. The expert judgment forecast was the worst (worse even than a naive forecast), and the ARIMA model led to the best price performance, followed closely by the composite forecast, a simple average of the econometric, ARIMA, and expert opinion forecast.

This article helps in comparing the models build by naive forecast, average forecast arima based on the AIC values, this was useful for me to select my model based for these methods.

CASE5: [5]

The use of time series modeling for the determination of rainfall climates of Iran S. Soltani,a R. Modarresa, * and S. S. Eslamian:

In this case the Time series modeling of the major cities of Iran was analyzed in this study. The Box–Jenkins popular ARIMA model was applied and seemed to fit the monthly rainfall time series very well.

Below are the details of the case

The process of time series modeling begins with the selection of the preliminary models interpreted from the characteristics of ACF and PACF functions using SAS ARIMA procedure (SAS/ETS, 1999). At first look, the monthly fluctuations show the seasonal behavior of the temporal pattern of the monthly rainfall due to the significant correlation coefficients at lag k = 12. For example, the ACF and PACF of the Ahwas, Isfahan and Ghaemshahr monthly rainfall series are presented in Figure 3. The parameter estimation of the preliminary selected models is then applied using the method of maximum likelihood. This model has been derived based on trying several models with different orders of the parameters. As the stationarity conditions are accepted for the model, the model residuals were checked for stationarity and normality using Portmanteau lack of fit test and normal tests. The portmanteau lack of fit test and the two normal tests (Kolmogrov–Smirnov and Anderson–Darling tests) proved the residuals to be time-independent (stationarity) and normally distributed. As a result from the above monthly rainfall time series modeling steps for Isfahan station, the best model for this station is ARIMA(1,0,0)(0,1,1)12. Plotting the observed and the model predicted rainfall time series shows that the model performs the observed rainfall series very well . Following the above procedures for all the selected stations, the best model for each station was estimated and presented in column 2 of Table III. In columns 3 and 4, the lag 1 and lag 12 autocorrelation coefficient values are also shown. It is clear from Figure that the model predicted rainfall series have the same seasonal fluctuations with the observed rainfall series. For better verification of the selected models and for checking their efficiency, two criteria are used, the correlation coefficient, R2, between observed and model predicted rainfall series and the R2 N –S criterion of Nash and Sutcliffe (1970). It is related to the sum of the squares of the differences, F, between the estimated and observed rainfall. This criterion is defined by R2 N –S = F° − F F° (5) where F° is the sum of the squares of differences between the observed rainfall and the mean rainfall. A value of R2 N –S greater than 90% would normally indicate a very satisfactory model performance while a value in the range 80–90% is regarded as an indication of a fairly good model. Values of R2 N –S in the range 60–80% generally indicate an unsatisfactory model fit (Shamseldin and O’Connor, 2001). The correlation coefficients and R2 criterion of Nash and Sutcliffe are presented in columns 5 and 6 of Table III, which postulate that the rainfall predicted

by the models fits correctly the observed values with R2 > 0.78 and R2 N –S > 85% to the observed rainfall and the fitted ARIMA models are satisfactory in all stations.

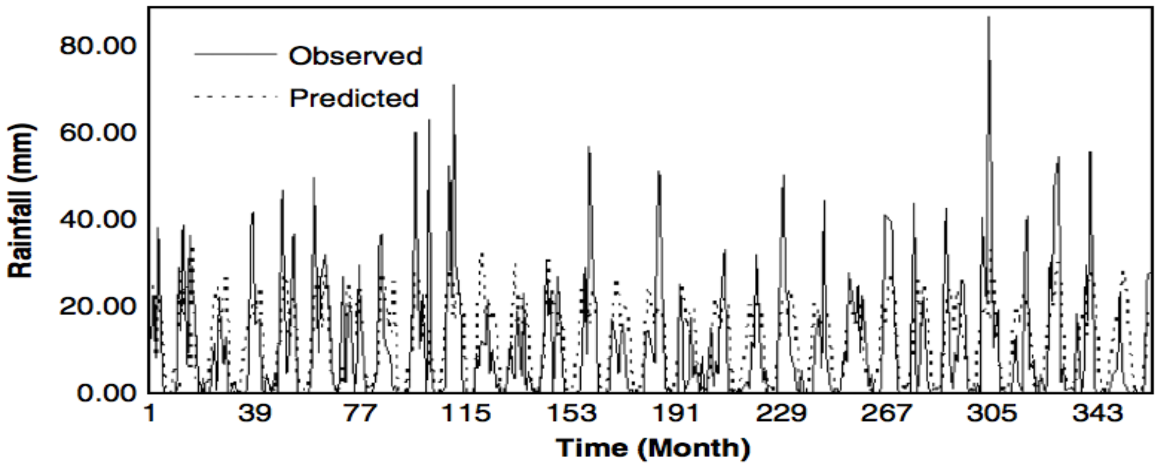


Figure 4. Time series of observed and model predicted rainfall for Isfahan station.

Table:3

Station (1)	Best model (2)	Values of autocorrelation coefficient		R^2 (5)	R^2_{N-S} (%) (6)	Groups (7)
		Lag-one (3)	Seasonal (lag-12) (4)			
Ahwaz	ARIMA(3,0,0) × $(0,1,1)_{12}$	0.27	0.36	0.92	90.22	3
Arak	ARIMA(1,0,0) × $(0,1,1)_{12}$	0.42	0.44	0.94	90.31	2
Ardabil	ARIMA(1,0,0) × $(7,1,1)_{12}$	0.17	0.2	0.89	86.04	3
Bandarabbas	$\text{ARIMA}(1,1,1)_{12}$	0.38	0.33	0.91	90.03	1
Bushehr	$\text{ARIMA}(1,1,1)_{12}$	0.35	0.3	0.91	90.25	1
Ghaemshahr	$\text{ARIMA}(0,1,1)_{12}$	0.18	0.28	0.90	89.6	1
Gorgan	$\text{ARIMA}(0,1,1)_{12}$	0.28	0.32	0.90	89.26	1
Ghazvin	ARIMA(0,0,1) × $(0,0,1)_{12}$	0.44	0.46	0.93	90.8	2
Hamedan	ARIMA(8,0,11) × $(19,1,1)_{12}$	0.47	0.55	0.87	85.09	3
Isfahan	ARIMA(1,0,0) × $(0,1,1)_{12}$	0.48	0.4	0.90	90.01	2
Ilam	$\text{ARIMA}(1,1,1)_{12}$	0.25	0.41	0.92	91.01	1
Oroumieh	ARIMA(6,0,0) × $(0,6,1)_{12}$	0.34	0.31	0.89	90.0	3
Ghom	ARIMA(1,0,4) × $(4,1,1)_{12}$	0.33	0.42	0.88	85.78	3
Zahedan	$\text{ARIMA}(1,0,1)_{12}$	0.27	0.17	0.94	91.1	1
Zanjan	ARIMA(6,0,0) × $(0,1,1)_{12}$	0.41	0.46	0.91	90.3	3
Yazd	$\text{ARIMA}(1,0,1)_{12}$	0.38	0.42	0.89	89.5	1
Yasuj	ARIMA(2,0,1) × $(0,1,1)_{12}$	0.32	0.33	0.90	90.06	2
Tehran	$\text{ARIMA}(1,0,1)_{12}$	0.38	0.4	0.91	90.7	1
Tabriz	ARIMA(0,0,1) × (1,0,1)	0.29	0.18	0.91	90.54	2
Shiraz	$\text{ARIMA}(1,0,1)_{12}$	0.28	0.24	0.92	90.87	1
Shahrecord	$\text{ARIMA}(1,0,1)_{12}$	0.23	0.36	0.90	88.6	1
Semnan	$\text{ARIMA}(1,1,1)_{12}$	0.2	0.35	0.93	91.5	1
Sanandaj	$\text{ARIMA}(0,1,1)_{12}$	0.21	0.24	0.94	92.61	1
Rasht	$\text{ARIMA}(0,1,1)_{12}$	0.41	0.46	0.94	93.1	1
Mashad	ARIMA(1,0,0) × $(1,1,1)_{12}$	0.25	0.24	0.89	86.4	2
Khoramabad	$\text{ARIMA}(1,0,1)_{12}$	0.53	0.51	0.92	89.74	1
Kermanshah	$\text{ARIMA}(1,0,1)_{12}$	0.21	0.29	0.91	90.2	1
Kerman	$\text{ARIMA}(1,1,1)_{12}$	0.32	0.34	0.91	90.25	1

4. Limitations and future work and Learning:

After verifying and validating the two models with various test and residual plots I could see that the models have a large scope of improving. Specially the regression models, I have tried with Year month as a covariate but it did not yield a good result. I ran the regression model against variables like t, t-1, t^2 but did not see any good results with the data. Also tried the model with sin.t and cos.t as independent variables but did not see any changes. I think the Adjusted R square value can be increased by transforming the data further and working with the outliers. For my current models I did not focus much on the outliers which would have definitely impacted the predictions.

As part of the future work I would definitely like to do some research work on the meteorological data and try to get the data for variables like temperature, density and further work on the model so see if the models can be optimized further. Also would like to work on the outliers to see if I can work on those and take appropriate actions. I would also like to work on my Arima model so that I can further optimize it build the models involving other parameters as well.

As far as learning goes I had never worked on time series data before, so this Was a good experience working on TS and also has a good experience working with forecasting techniques including the simple techniques such as average method, naïve and drift method, apart from the techniques the things that I have learned while formulating the model is understanding the data and exploratory analysis and research regarding the data really helps in the process of building the models, also deploying the model in such a way the models are re-usable is important aspect that I have learned from this process. And during the process of building the models I have come across various techniques in R to work with data frames/time series and also to create functions which can be used to call and run the models whenever required.

5. References:

[1] http://www.ijcse.com/docs/INDJCSE14-05-02-081.pdf
[2] http://eccsf.ulbsibiu.ro/articole/vol91/918kumar&anand.pdf
[3] http://wwwpersonal.umich.edu/~copyrght/image/solstice/sum07/Solstice_GoutamiED.pdf
[4] https://faculty.fuqua.duke.edu/~clemen/bio/Published%20Papers/13.CombiningReview-Clemen-IJOF-89.pdf
[5] http://onlinelibrary.wiley.com/store/10.1002/joc.1427/asset/1427_ftp.pdf?v=1&t=iuivvg8y&s=08d31600e59a1e4b8f4affebf043004a6683556f
[6] Jyouti Upadhaya,Assam University .” Climate Change and its impact on Rice productivity in Assam” .
[7] Olaiya Folorunsho(2012):Application of Data mining Techniques in Weather Prediction and Climate change studies
[8] M.Kannan;S.Prabhakaran;P.Ramchandran- “Rainfall forecasting using DM Technique-IJET 10-2-06-28.
[9] Reyson P.Raymundo- “Rainfall Forecasting Model in the province of Isabela .IMURE:International Journal of Mathematics,Engg Technology.

[10]Coghlan, Avril (2010), A Little Book of R for Time Series, Readthedocs.org, Available online at: http://a-little-book-of-r-for-time-series.readthedocs.org/en/latest/src/timeseries.html

[11] Contreras, J., Espinola, R., Nogales, F.J., Conejo, A.J., (2003), *ARIMA Models to Predict Next- day Electricity Prices*, IEEE Transactions on Power Systems, Vol. 18, No. 3, pp. 1014- 1020.

[12] Hannan, E., (1980), *The Estimation of the Order of ARMA Process*, Annals of Statistics, Vol. 8, pp. 1071-1081.

[13] https://kaggle2.blob.core.windows.net/competitions-data/inclass/5548/GHCND_documentation.pdf?sv=2012-02-12&se=2016-10-26T18%3A31%3A48Z&sr=b&sp=r&sig=BJDO%2Fb%2FKbnvWD7HElivtR8V%2FG9cxuvMG15IQYySD9aI%3D
[14] http://www.dataiku.com/learn/guide/code/r/time_series.html