Prediction Analysis through time series data

Project Report
by

Pavan Kumar Reddy Gunnala,811173384, pgunnala@kent.edu

Under the guidance of
Dr Ruoming Jin
Professor

From
Kent State University

1. Introduction

Motivation examples of this project

Time series classification is a procedure that involves sorting chronological data into categories. It's employed in a variety of fields, including meteorology, medicine, and physics. Many algorithms have been developed in the previous decade to execute this task with high accuracy. However, applications with uncertain time series have been under-explored. We present a new uncertain dissimilarity measure based on uncertainty propagation techniques. The huge tests we ran using state-of-the-art rainfall datasets demonstrate the value of our contribution.

Real applications

Covid-19 Cases Predictions for Next 30 Days
Stock Price Prediction using Linear Regression
Earthquake Prediction Model
Predict Migration
Weather Prediction Model
Time Series with LSTM
Daily Births Forecasting
Google Stock Price Prediction
Anomaly Detection using ARIMA Model
Rainfall Prediction Model

2. Project Description

Brief descriptions of your project

Rainfall data are a crucial input for various tasks concerning the wet weather period. Nevertheless, their measurement is affected by random and systematic errors that cause an underestimation of the

rainfall volume. Therefore, the general objective of the presented work was to assess the credibility of measured rainfall data. Studies of the hydroclimate at regional scales rely on spatial rainfall data products, derived from remotely sensed (RS) and *in-situ* (IS, rain gauge) observations. Because regional rainfall cannot be directly measured, spatial data products are biased. These biases pose a source of uncertainty in environmental analyses, attributable to the choices made by data-users in selecting a representation of rainfall. In order to quantify total error affecting hydrological models and predictions, we must explicitly recognize errors in input data, model structure, model parameters and validation data. This tackles the last of these: errors in discharge measurements used to calibrate a rainfall-runoff model, caused by stage–discharge rating-curve uncertainty. This uncertainty may be due to several combined sources, including errors in stage and velocity measurements during individual gauging, assumptions regarding a particular form of stage–discharge relationship, extrapolation of the stage–discharge relationship beyond the maximum gauging, and cross-section change due to vegetation growth and/or bed movement. So, we will be predicting the future outcome using the old data set.

In recent days uncertain data is creating a hindrance in big data which results in no proper estimations of rainfall, stocks, and many others. So, we deal with the probabilistic uncertain data of a time series dataset, where the uncertainties may cause due to sensory defects, missing data or a large fluctuation in data due to any external factors. A forecast is calculation or estimation of future events, especially for financial trends or coming weather. Until this year, forecasting was very helpful as a foundation to create any action or policy before facing any events. As an example, in the tropics region which several countries only had two seasons in a year (dry season and rainy season), many countries especially country which relies so much on agricultural commodities will need to forecast rainfall in term to decide the best time to start planting their products and

maximizing their harvest. Another example is forecast can be used for a company to predict raw material prices movements and arrange the best strategy to maximize profit from it.

In this project, the main process we follow during prediction is to identify the uncertainties of time series data and overcome come them using various techniques. Later we will be training and testing the data using pattern matching technique. The model we are going to use in this project is ARIMA (Auto Regressive Integrated Moving Average) which leads us to gain a confidence level by comparing trained data with the testing data.

Challenges in our project

Data Problems to Consider, here are some of the challenges facing

Incorrect data
Missing data
Insane Amounts of Data
Poor Data Quality
Data Accessibility

These are just the data challenges we face BEFORE beginning the actual work of trying to solve problems with data. First real endeavors working through the big data mess will encourage it to create data governance and to support stronger data curation practices across all industries.

A major challenge in climate prediction is the uncertainty on how we are going to deal with climate change. Computer simulations must be run again and again with different scenarios that vary in future economic development, amounts of climate- influencing gases, the change in use of land use practices, political decisions, etc.

3. Background

- Related papers (or surveys for graduate teams)

Many decades ago, in a galaxy not too far away from here, statisticians analyzed time series data without taking into account how ‘no stationariness’(read: growth/decline over time) might have an effect on their analyses. Then George P. Box and Gwilym Jenkins came along and presented a famous monograph called “Time Series Analysis: Forecasting and Control” in which they showed that nonstationary data could be made stationary (read: steady over time) by “differencing” the series. In this way, they could pull apart a juicy trend at a specific time period from a growth/decline that would be expected anyway, given the no stationariness of the data.

Methodology and implementation details of ARIMA model-based seasonal adjustment are presented, with key features illustrated with the simplest formulas that provide concrete representations of the method, which was developed by Tiao and Hillmer (1978) and Hillmer and Tiao (1982), with important implementation contributions by Burman (1980), Gomez and Maravall (1996), and others.

Software tools

R Language
R Studio
Libraries Required:
"reshape","forecast","tseries","rmarkdown","knitr","ggplot2"
Dataset: Dataset with uncertainties

Required hardware

System Specification: Computer with I3 processor ,4 GB RAM
Programming Skills: R language

Related programming skills (functions, Internet programming, object-oriented programming, distributed environment, etc.)

In this project we have also applied differ methods like Mean, Naïve and seasonal Naïve to compare the results with our Model. So having the prior idea on those methods also helps in this project.

4. Problem Definition

Formal (mathematical) definitions of problems

Definition 1 (Time series):
A time series T of length m is an ordered sequence of m observations t_i .
T =< t1, t2, ..., tm >

Definition 2 (Uncertain time series):
An uncertain time series (UTS) is a time series in which the observations are uncertain. Uncertain observations can be modeled with the multiset based or the probability density function-based (PDF-based) modeling:

In this section, we describe how to classify uncertain time series using shapelets. Uncertain observations are represented using the probability density function model (or simply pdf model). We start by defining the concepts that are used in our algorithm, then we describe the algorithm itself.

Definition 3 (Uncertain subsequence): An uncertain subsequence S of an uncertain time series T is a series of l (its length) consecutive uncertain values in T.

$$S = S\hat{} \pm \delta S = \{ti+1 \pm \delta t_{i+1}, ..., t_{i+l} \pm \delta t_{i+l}\}$$
$$1 \leq i \leq m - l, 1 \leq l \leq m, m = |T$$

Definition 4 (Uncertain dissimilarity): The dissimilarity between two uncertain subsequences S and R is the uncertain distance between them

$$d = UED(S, R) = UED(R, S)$$

The dissimilarity between an uncertain time series T and an uncertain subsequence S is the dissimilarity between S and the subsequence of T that is the most similar to S. It is formally defined as follows:

$$UED(T, S) = \min\{UED(S, R) \mid \forall R \subseteq T, |S| = |R|\}$$

Definition 5 (Uncertain separator): An uncertain separator sp for a dataset D of uncertain time series is an uncertain subsequence that divides D in two parts D1 and D2 such that:

$$D1 = \{T \mid UED(T, sp) \leq €, \forall T \in D\}$$
$$D2 = \{T \mid UED(T, sp) > €, \forall T \in D\}$$

As in Definition 4, the quality of a separator is measured using the information gain (IG). Given the previous definitions, we can give a formal definition of an uncertain shapelet.

Definition 6 (Uncertain shapelet): An uncertain shapelet S for a dataset D of uncertain time series is an uncertain separator that maximized the information gain.

$$S = \text{argmax}_{sp}(IG(D, sp))$$

Challenges of tackling the problems

Time series classification problems are differentiated from traditional classification problems because the attributes are ordered and long. Whether the ordering is by time or not, it is in fact related with each other. The important characteristic is that there important features dependent on the ordering. Classification, which is the task of assigning objects to one of several predefined categories, which is a pervasive problem that encompasses many diverse applications.

Definition (Classification).Classification is the task of learning a target function that maps each attribute set x to one of the predefined class labels y. The target function is knowns informally as a classification model. A case is a pair{x,y} with m observations $x1 , \dots , xm$ (the time

series) and discrete class variable y, which is c possible values. n represents the number of sample points.

$$T =< X, Y >=< (x_1,y_1), \dots , (x_n,y_n) >$$

A brief summary of general solutions in your project

Uncertain shapelet transform classification

Our Project for uncertain time series classification is an extension of the shapelet transform algorithm. Given a dataset D of uncertain time series, the first step is to select the top k best uncertain shapelets from the dataset. This step is achieved using the procedure described, which takes as input, the dataset D, the maximum number of uncertain shapelets to be extracted k, the minimum and the maximum length of an uncertain shapelet MIN and MAX. This algorithm uses three subprocedures:

• GenCand(T, MIN, MAX) which generates every possible uncertain shapelet candidates from the input uncertain time series T. These candidates are uncertain subsequences of T, with length at least MIN and at most MAX.

• AssessCand(cands, D) which computes the quality of each candidate in the list of candidates cands. The quality of a candidate is the information gain it produces when used as a separator for the dataset.

• ExtracBest(C, Q, k) which takes the list of uncertain shapelet candidates C, their associated qualities Q and returns first k uncertain shapelets with highest qualities

In summary, This generates every uncertain subsequences of length at least MIN and at most MAX from the dataset, assesses the quality of each one by computing the information gain obtained when it is used as a separator for the dataset and finally returns the k subsequences that produce the highest information gain. The parameters MIN and MAX should be optimized to reduce the execution time of the algorithm. With

the knowledge of the domain, the length of a typical shapelet can be estimated and used to set MIN and MAX in order to reduce the number of shapelet candidates. By default MIN is set to 3 and MAX is set to m − 1, where m is the length of the time series.

function USHAPELETSELECTION(D, k, MIN, MAX)
  $C \leftarrow \emptyset; Q \leftarrow \emptyset$
  for $i \leftarrow 1, n$ **do**
    $cands \leftarrow GenCand(T_i, MIN, MAX)$
    $qualities \leftarrow AssessCand(cands, D)$
    $C \leftarrow C + cands$
    $Q \leftarrow Q + qualities$
  end for
  $S \leftarrow ExtractBest(C, Q, k)$
  return S ▷ Top k uncertain shapelets
end function

The next step after the top-k uncertain shapelets selection is the uncertain shapelet transformation. This step is that, which takes as input the dataset D, the set of the top-k uncertain shapelets S and the number of uncertain shapelets k. For each uncertain time series in the dataset, its uncertain feature vector of length k is computed using UED. The i th element of the vector is the UED between the uncertain time series and the uncertain shapelet i. Because the uncertainties add up during the transformation, the uncertain feature vectors are such that the scale of the best guesses is smaller than the scale of the uncertainties. It is very important to have everything on the same scale.

function USHAPELETTRANSFORMATION(D, S, k)
  for $i \leftarrow 1, n$ **do**
    $temp \leftarrow \emptyset$
    for $j \leftarrow 1, k$ **do**
      $temp_j \leftarrow UED(T_i, S_j)$
    end for
    $D_i \leftarrow temp_j$
  end for
  for $i \leftarrow 1, n$ **do**
    for $j \leftarrow 1, k$ **do**
      $best \leftarrow \frac{\hat{D_{ij}} - mean(\hat{D_{:j}})}{std(\hat{D_{:j}})}$
      $delta \leftarrow \frac{\delta D_{ij} - mean(\delta D_{:j})}{std(\delta D_{:j})}$
      $D_{ij} = best \pm delta$
    end for
  end for
  return D ▷ The transformed dataset
end function

The third and last step is the effective classification. A supervised classifier is trained on the uncertain transformed dataset, such that, given the feature vector of an unseen uncertain time series, it can predict its class label. Since the uncertainty have been propagated, the training process can be aware of uncertainty by taking it as part of the input. More specifically, best guesses are features and uncertainties are features of best guesses, and thus are metafeatures. There exists many supervised classifiers in the literature for the classification of uncertain tabular data.

Now let's say we have a temporal series of binary targets and we try to predict them fitting an ARIMA. In this process, we deal with a series of monthly changes in Rainfall data. The changes of each month are obtained as the ratio between the data of this month (Mt) and the previous month (Mt-1) . A ratio higher than 1 means that there is an increase in the rainfall for that month concerning the previous one. Our binary targets are 1 for positive changes and 0 for negative changes.

Given a series of binary targets, we transform them into soft labels and apply a logit function. We fit an ARIMA selecting the best hyperparameters. The final predictions are obtained applying the sigmoid transformation. In the plot below, we show the true labels , in the background, and the predicted probabilities for the train set and the test set.

The results are very interesting. Our approach produces good predictions showing the ability to produce outcomes in a probability format. The same can be generalized to every binary time series or extended to multiclassification as a one-vs-rest procedure.

In this, we introduced a technique to carry out classification tasks with soft labels and regression models. Firstly, we applied it with tabular data,

and then we used it to model time-series with ARIMA. Generally, it is applicable in every context and every scenario, providing also probability scores. This may be extremely useful also for label smoothing to reduces some mistakes present in labeling.

5. The Proposed Techniques

- Framework (problem settings)

  The framework we used in this project is R studio with R language installed in it. The necessary packages we need to include while computing this project are "reshape","forecast","tseries","rmarkdown","knitr","ggplot2", So that we'll be able to run the necessary scripts to compute the dataset with our modelling.

- Details of major techniques

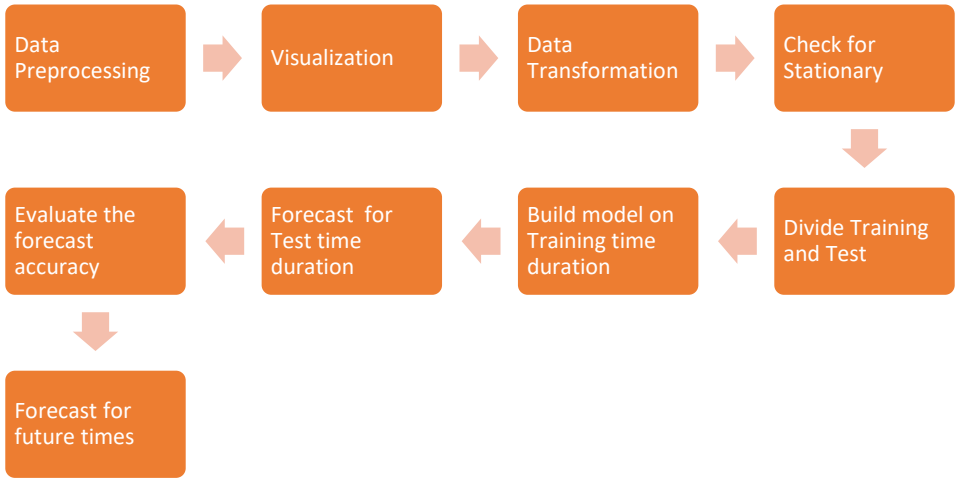
  The major technique we used to predict the future outcome is by using ARIMA model where the uncertain data is classified with the shaplet transform algorithm where all the classification definitions are explained above.

  Auto regressive integrated moving average model ARIMA model is a class of statistical models for analyzing and forecasting time series data. It explicitly caters to a suite of standard structures in time series data, and provides a simple powerful method for making skillful time series forecasts[5].This acronym is descriptive, capturing the key aspects of the model itself. Briefly, they are: AR: Autoregression. A model that uses the dependent relationship between an observation and some number of lagged observations. I: Integrated. The use of differencing of raw observations (e.g. subtracting an observation from an observation at the previous time step) in order to make the time series stationary. MA: Moving Average. A model that uses the

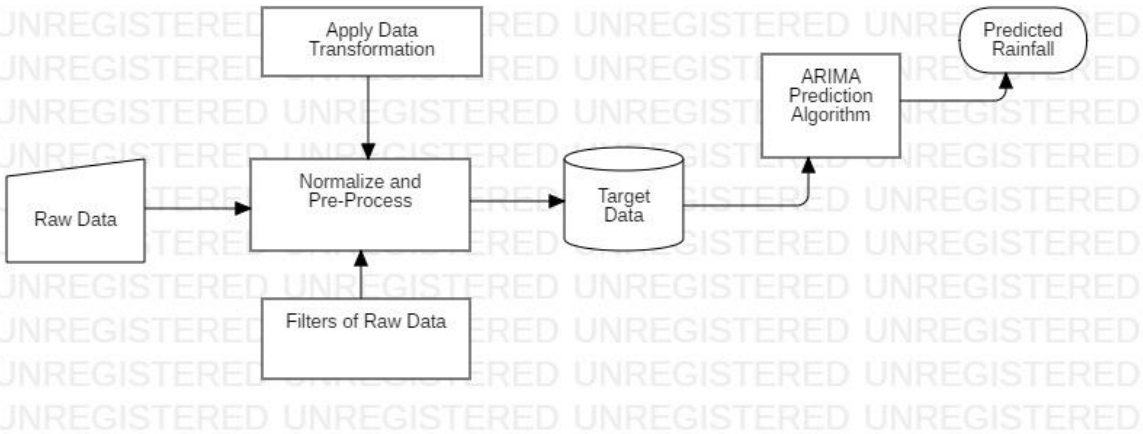
dependency between an observation and a residual error from a moving average model applied to lagged observations.

6. Visual Applications

- Design modules (with descriptions, figures, and/or flowcharts)



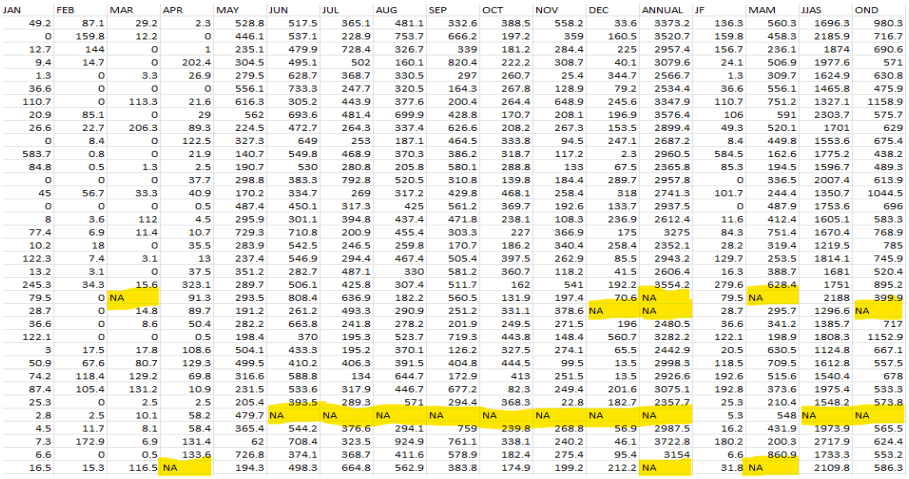
The above figure shows the flow which we followed in this project all the data shapelets are done in the preprocessing phase.



This diagram shows about the Design flow where we took the raw data with uncertainties and applied few techniques as explained above to remove uncertainties from the dataset and then after classifying uncertainties we have applied the ARIMA Prediction algorithm to predict the future time series data using the historic dataset.

7. Experimental Evaluation

- Experimental settings
  - Descriptions of real/synthetic data sets



JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC	ANNUAL	JF	MAM	JJAS	OND
49.2	87.1	29.2	2.3	528.8	517.5	365.1	481.1	332.6	388.5	558.2	33.6	3373.2	136.3	560.3	1696.3	980.3
0	159.8	12.2	0	446.1	537.1	228.9	753.7	666.2	197.2	359	160.5	3520.7	159.8	458.3	2185.9	716.7
12.7	144	0	1	235.1	479.9	728.4	326.7	339	181.2	284.4	225	2957.4	156.7	236.1	1874	690.6
9.4	14.7	0	202.4	304.5	495.1	502	160.1	820.4	222.2	308.7	40.1	3079.6	24.1	506.9	1977.6	571
1.3	0	3.3	26.9	279.5	628.7	368.7	330.5	297	260.7	25.4	344.7	2566.7	1.3	309.7	1624.9	630.8
36.6	0	0	0	556.1	733.3	247.7	320.5	164.3	267.8	128.9	79.2	2534.4	36.6	556.1	1465.8	475.9
110.7	0	113.3	21.6	616.3	305.2	443.9	377.6	200.4	264.4	648.9	245.6	3347.9	110.7	751.2	1327.1	1158.9
20.9	85.1	0	29	562	693.6	481.4	699.9	428.8	170.7	208.1	196.9	3576.4	106	591	2303.7	575.7
26.6	22.7	206.3	89.3	224.5	472.7	264.3	337.4	626.6	208.2	267.3	153.5	2899.4	49.3	520.1	1701	629
0	8.4	0	122.5	327.3	649	253	187.1	464.5	333.8	94.5	247.1	2687.2	8.4	449.8	1553.6	675.4
583.7	0.8	0	21.9	140.7	549.8	468.9	370.3	386.2	318.7	117.2	2.3	2960.5	584.5	162.6	1775.2	438.2
84.8	0.5	1.3	2.5	190.7	530	280.8	205.8	580.1	288.8	133	67.5	2365.8	85.3	194.5	1596.7	489.3
0	0	0	37.7	298.8	383.3	792.8	520.5	310.8	139.8	184.4	289.7	2957.8	0	336.5	2007.4	613.9
45	56.7	33.3	40.9	170.2	334.7	269	317.2	429.8	468.1	258.4	318	2741.3	101.7	244.4	1350.7	1044.5
0	0	0	0.5	487.4	450.1	317.3	425	561.2	369.7	192.6	133.7	2937.5	0	487.9	1753.6	696
8	3.6	112	4.5	295.9	301.1	394.8	437.4	471.8	238.1	108.3	236.9	2612.4	11.6	412.4	1605.1	583.3
77.4	6.9	11.4	10.7	729.3	710.8	200.9	455.4	303.3	227	366.9	175	3275	84.3	751.4	1670.4	768.9
10.2	18	0	35.5	283.9	542.5	246.5	259.8	170.7	186.2	340.4	258.4	2352.1	28.2	319.4	1219.5	785
122.3	7.4	3.1	13	237.4	546.9	294.4	467.4	505.4	397.5	262.9	85.5	2943.2	129.7	253.5	1814.1	745.9
13.2	3.1	0	37.5	351.2	282.7	487.1	330	581.2	360.7	118.2	41.5	2606.4	16.3	388.7	1681	520.4
245.3	34.3	15.6	323.1	289.7	506.1	425.8	307.4	511.7	162	541	192.2	3554.2	279.6	628.4	1751	895.2
79.5	0	NA	91.3	293.5	808.4	636.9	182.2	560.5	131.9	197.4	70.6	NA	79.5	NA	2188	399.9
28.7	0	14.8	89.7	191.2	261.2	493.3	290.9	251.2	331.1	378.6	NA	NA	28.7	295.7	1296.6	NA
36.6	0	8.6	50.4	282.2	663.8	241.8	278.2	201.9	249.5	271.5	196	2480.5	36.6	341.2	1385.7	717
122.1	0	0	0.5	198.4	370	195.3	523.7	719.3	443.8	148.4	560.7	3282.2	122.1	198.9	1808.3	1152.9
3	17.5	17.8	108.6	504.1	433.3	195.2	370.1	126.2	327.5	274.1	65.5	2442.9	20.5	630.5	1124.8	667.1
50.9	67.6	80.7	129.3	499.5	410.2	406.3	391.5	404.8	444.5	99.5	13.5	2998.3	118.5	709.5	1612.8	557.5
74.2	118.4	129.2	69.8	316.6	588.8	134	644.7	172.9	413	251.5	13.5	2926.6	192.6	515.6	1540.4	678
87.4	105.4	131.2	10.9	231.5	533.6	317.9	446.7	677.2	82.3	249.4	201.6	3075.1	192.8	373.6	1975.4	533.3
25.3	0	2.5	2.5	205.4	393.5	289.3	571	294.4	368.3	22.8	182.7	2357.7	25.3	210.4	1548.2	573.8
2.8	2.5	10.1	58.2	479.7	NA	NA	NA	NA	NA	NA	NA	NA	5.3	548	NA	NA
4.5	11.7	8.1	58.4	365.4	544.2	376.6	294.1	759	239.8	268.8	56.9	2987.5	16.2	431.9	1973.9	565.5
7.3	172.9	6.9	131.4	62	708.4	323.5	924.9	761.1	338.1	240.2	46.1	3722.8	180.2	200.3	2717.9	624.4
6.6	0	0.5	133.6	726.8	374.1	368.7	411.6	578.9	182.4	275.4	95.4	3154	6.6	860.9	1733.3	553.2
16.5	15.3	116.5	NA	194.3	498.3	664.8	562.9	383.8	174.9	199.2	212.2	NA	31.8	NA	2109.8	586.3

The dataset we took in this project is a real time dataset which consists of different states data from the years 1901 to 2018. The dataset consists of uncertainties as shown in the above picture where we have the uncertainties like missing values, zero values and improper measurements.

  - Competitors (baseline method, or existing techniques to compare with)

    The baseline method we get from the internet sources is done by apply a existing model to analyze, test and train dataset. Here compared to the existing method we took the uncertain data from data.gov and worked on the uncertainties found in that later we applied the modelling technique this made us different from all previous internet-based baseline methods.
  - Parameter settings

    In the dataset we used in this dataset we have the different attributes of rainfall like months, annual and seasonal data of different states of India. So, we declared the parameters as a state's data, decomposition and evaluating results.
  - Evaluation measures

    For evaluating the prediction, we used MAPE Function

The mathematical formula to calculate MAPE is:

$$MAPE = (1/n) * \Sigma(|Original – Predicted| / |Original|) * 100$$

where:

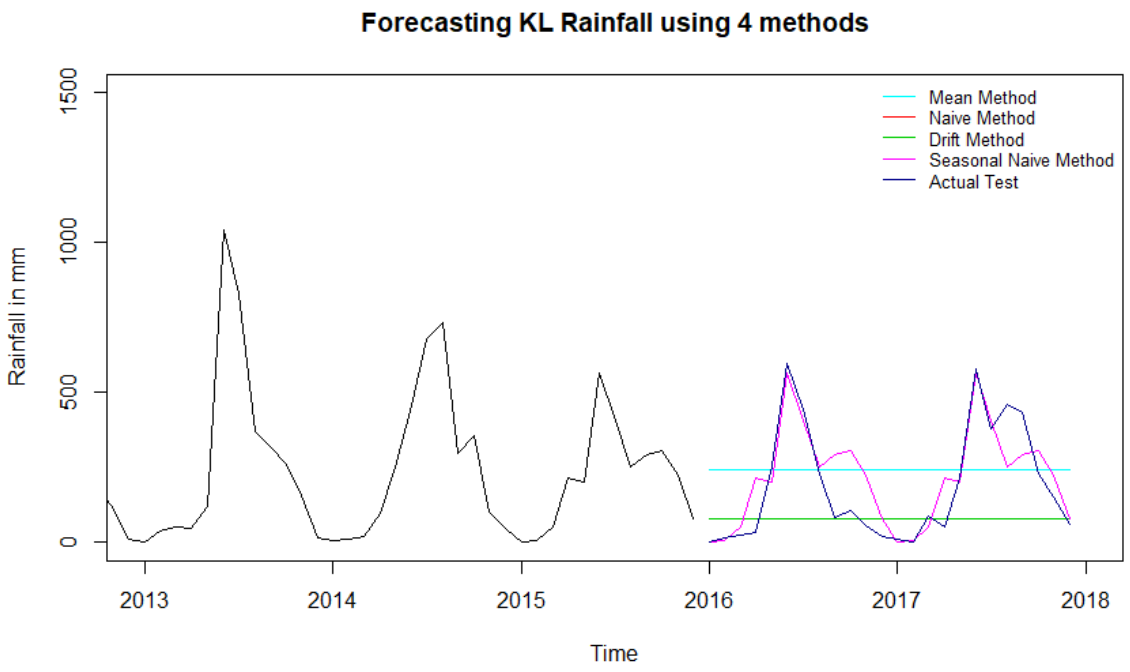
Σ –indicates the “sum”

n – indicates the sample size

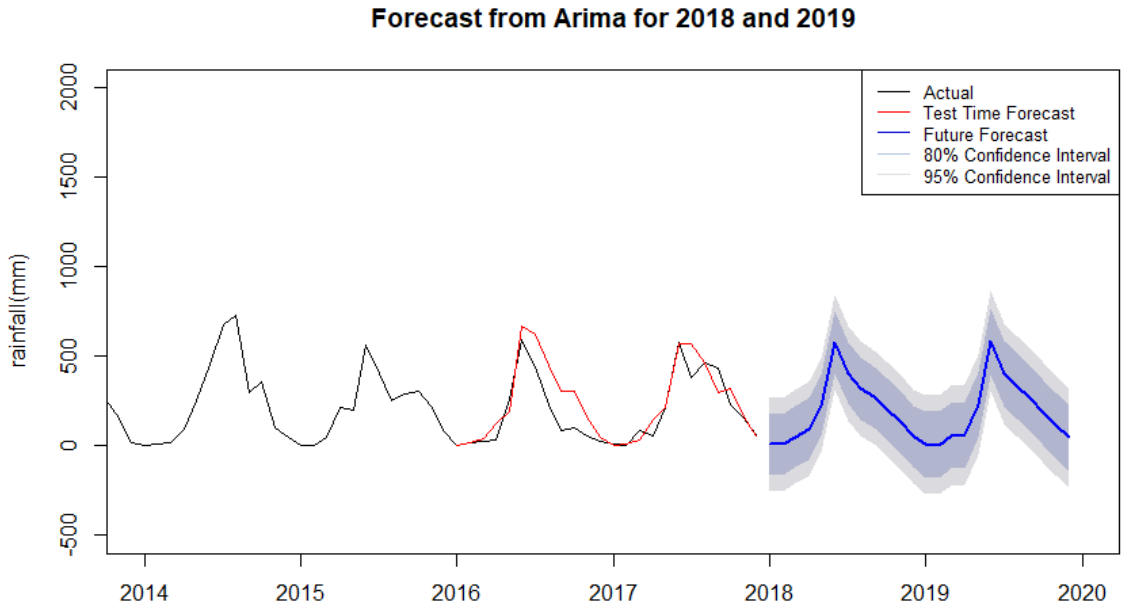
actual – indicates the actual data value

forecast – indicates the forecasted data value

- The performance report (pruning power, recall/precision/f-measure, CPU time, I/O cost, communication cost, index construction time/space, etc.)

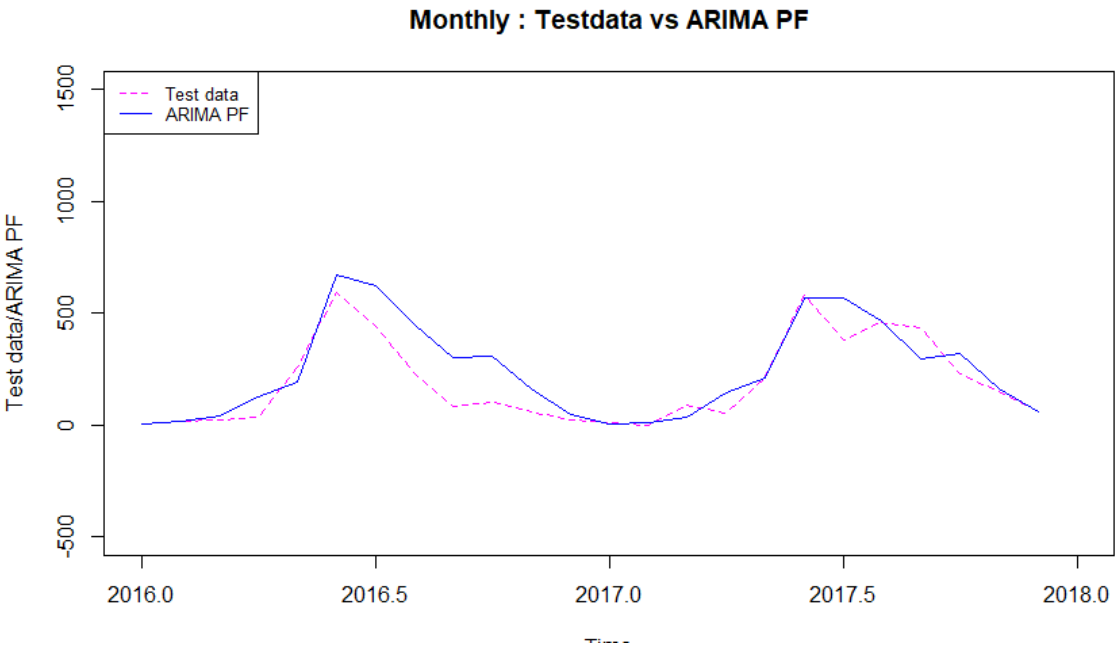


Her we have compared the performance with Mean, Naïve, Drift and seasonal Naïve Methods to compare it with our model so this above graphical representation shown this our model as got the greater confidence level compared to other models.



The performance we got from this project is shown in the above graphical presentation.

- Screen captures



The prediction we got from the final is so similar to the test data.

8. Future Work

Our project gives us the great confidence level of prediction compared to other model algorithms but we can deal the uncertainties in better methods other than our method, so in future we will be focusing in removal of uncertainties in better methods.

9. References

I. Ansari, H. (2013). Forecasting seasonal and annual rainfall based on non-linear modelling with Gamma test in North of Iran, International Journal of Engineering Practical Research, 2 (1) 16- 29.

II. Abudu, S., Cui, C.L., King, J.P., Abudukadeer, K. (2010). Comparison of performance of statistical models in forecasting monthly stream flow of Kizil River, China. Water Science and Engineering, 3(3) 269–281. Babu, S.K.K., Karthikeyan, K., Ramanaiah, M.V., Ramanah, D. (2011). Prediction of rain-fall flow time series using Auto-Regressive Models. Advances in Applied Science Research, 2(2) 128–133.

III. Chattopadhyay, S., Chattopadhyay, G. (2010). Univariate modelling of summer-monsoon rainfall time series: Comparison between ARIMA and ARNN. Comptes Rendus Geoscience, 342(2) 100–107. Collischonn, W., Haas, R., Andreolli, I., and Tucci, C.E.M. (2005).

IV. Forecasting river Uruguay flow using rainfall forecasting from a reagional weather predicted model, Journal of hydrology, 305 (1-4) 87-98.

V. Engelbert Mephu Nguifo, Michael Franklin Mbouopda- Uncertain Time Series Classification With Shapelet Transform- February 2021