FEATURE EXTRACTION TO IDENTIFY PREDICTORS FOR TIME SERIES ANALYSIS OF INDIAN SUMMER MONSOON RAINFALL

Kaushlendra Yadav*

Shraddha Srivastava*

ABSTRACT

An accurate forecast of seasonal summer monsoon rainfall over the country is an increasing demand for decision makers and planners of the country in mitigating any kind of disaster like food crisis and water scarcity.

The aim is to identify predictors for the Indian summer monsoon rainfall, which can then be used in forecast models. The parameters which show high correlation coefficient are identified as the predictors.

It is proposed to extend the study carried out by Shraddha Srivastava and K.C. Tripathi (2012): wherein it was concluded that a better approach for deciding the predictors would be to do the PCA (principal component analysis)for the predictors and then making a better ANN architecture or regression equation for the forecasting purpose

General Terms

Artificial neural network

Keywords

PCA (principal component analysis), Indian summer monsoon rainfall (ISMR)

* M.Tech (CSE), Inderprastha engg. College, UPTU LUCKNOW

1. INTRODUCTION

If we want to predict anything then we need certain data as input and a output by which we can do analysis that how much accurate we are in our prediction .we can do good prediction if our trained output values are closer to the known output values. So here we will use past data of Indian summer monsoon rainfall for future prediction. Our prediction basically can be done by seeing the behaviour of atmosphere around the ocean, on analysing many things like temperature of sea surface, by looking temperature of surface air and pressure of sea level etc.

Identification of Predictors

The aim is to identify predictors for the Indian summer monsoon rainfall, which can then be used in forecast models. The parameters which show high correlation coefficient are identified as the predictors. The number of predictors is different for different seasons.

- The two main requirements for any useful predictors are:
- 1) It should show a good variable property with respect to the seasonal summer monsoon rainfall. that will show how things are happening with respect to each other in a pattern way
- 2) How certain months are playing a lead role in a particular season.

.

2. Forecast Models

- It is proposed to carry out the prediction of the ISMR using the leading principal components of the predictors determined by the correlation analysis
- The prediction models proposed for the study are
- Regression model
- Artificial neural network model
- Principle component analysis model.
- The primary aim of all these models is to found the best correlation by which we easily understand the behavior of certain data, and find the pattern in data that will lead us to easily do prediction of ISMR.

- As we can easily understand that in a particular season few months plays key role than the other months, so while choosing predictors our main concern will be on founding the high correlation between the values that awe are considering as a predictor.

2.1 Regression model

In it we see the relationship between two (or more) variables by applying a mathematical formula.

if we express x as the predictor (independent) variable and y as the response (dependent) variable

Then we specifically want to indicate how y varies as a function of x.

In regression model we see that RMSE values are always less than the standard deviation of the observed data it means performance or regression model is fairly good enough as predictions are better than mean prediction. However performance of regression model is inferior to ANN model.

2.2 Artificial Neural Network

- The ANN model is based on 'prediction' by smartly 'analyzing' the trend from an already existing voluminous historical set of data. The other models are either mathematical or statistical. These models have been found to be very accurate in calculation, but not in prediction as they cannot adapt to the irregularly varying patterns of data which can neither be written in form of a function, nor deduced from a formula.
- These real-life situations have been found to be better interpreted by 'artificial neurons' which can learn from experience, i.e by back-propagation of errors in next guess and so on. This may lead to a compromise in accuracy, but give us a better advantage in 'understanding the problem', duplicating it or deriving conclusions from it.

.

2.3 Principal Component Analysis

- This model reconstructs the assumed deterministic dynamics of the monsoon rainfall data .We will further use regional and global parameters as the predictors which are chosen by examining their physical linkage with the monsoon and their degree of relationship with the monsoon rainfall of India. Since some of these predictors are intercorrelated, we will do principal component analysis of these parameters and few principal components will be taken as inputs to develop a feature vector.

  Steps followed in principal component analysis.

- Step 1: Get some data
- Step 2: Subtract the mean
- Step 3: Calculate the covariance matrix
- Step 4: after getting covariance matrix find the eigenvectors and eigenvalues of the covariance matrix.
- Step 5: find components and make a feature vector
- Step 6: Deriving the new data set

3. METHODOLOGY

- The methodology essentially involves time series prediction. We will design Back propagation ANN model with delta learning rule for the prediction of the all India rainfall. The most common and most useful statistics in our analysis is 'correlation'. When we are looking relationship between two variables then we will find that correlation is a value that describes at what extent, variables are related to each. In MATLAB correlation is found with the function "corrcoef(var1,var2)". In our study relation have been calculated taking 1lag to 12lag between the months of AIR data points of 140 years.

Actual data

First we will arrange our 140 years data of Indian summer monsoon years taking on one axis and corresponding month's data points on other axis.

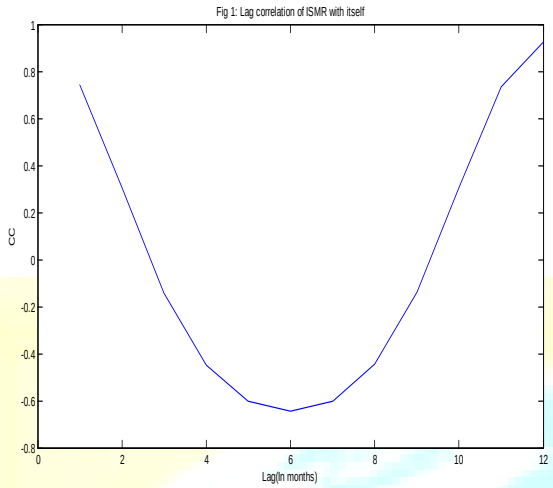
jan	feb	march	april	may	june	july	august	sept	octo	novemb	december
196	107	144	339	636	2080	2778	1794	1835	368	323	67
76	75	73	240	438	1892	2913	2451	1879	785	276	191
36	135	150	243	428	1130	2644	2142	1656	607	115	89
86	158	106	169	683	2278	3069	2334	2062	932	187	40
98	113	130	232	505	1926	3079	2186	2105	566	63	71
9	21	159	165	425	1232	2968	1956	1620	464	95	25
291	234	240	353	674	1423	1564	1569	1483	1077	188	364
108	91	102	361	665	1303	2940	3393	2124	798	273	140
21	107	80	82	877	1900	2241	3220	1616	865	205	71
38	165	151	211	495	1876	2717	1768	1840	877	528	111
13	32	330	236	530	1606	2927	2646	1433	493	347	61
100	94	90	188	596	2130	3311	1902	1685	739	596	61
179	36	189	165	621	2043	2690	1879	1880	830	282	155
70	74	82	203	383	1593	2924	2426	2385	866	269	304
96	98	136	199	528	1940	2728	2479	1327	737	321	501
48	12	222	124	711	1977	3159	2218	1390	1319	214	170
175	6	148	209	551	1898	3013	2574	1507	879	411	130
197	136	119	259	531	1323	2742	2800	1249	382	530	72
106	134	66	243	448	2035	2797	2780	1693	940	253	68
26	11	172	261	358	2279	3009	2141	1626	647	351	136
114	134	348	207	595	833	2571	2297	2222	602	172	71
53	106	56	328	541	1579	3138	3062	2134	993	126	69
224	283	467	252	946	2416	2568	2309	2256	969	642	22
114	144	181	247	359	2168	3114	2415	2017	1388	385	137

Lag data

Here we will see our data will be arranged when taking the leg relation among the months .if we are taking 1 lag that mean we r calculating February data points from the previews data points of January .we will continue in the same manner for lag 1 to lag 12 .

months	1lag	2lag
jan	feb	march
feb	march	april
march	april	may
april	may	june
may	june	july
june	july	august
july	august	september
august	september	october
september	october	november
october	november	december
november	december	jan
december	jan	feb
jan	feb	march
feb	march	april
march	april	may
april	may	june
may	june	july
june	july	august
july	august	september
august	september	october
september	october	november
october	november	december
november	december	
december		

Correlation graph



Correlation analysis

Autocorrelation value of a series with a lagged series of itself gives an insight of the level of dependence of the future values in the series on the present value. Statistical prediction models predict the time series ahead of time by exploiting the patterns in the past data.

It is seen that correlation have been found more than 0.5 for lag 1, lag 11, lag 12.

So we will use these three as predictors for ISMR.

4. Results and Discussion

We found good correlation at lag 1, lag11, lag12.

Now we will compare these lag data with lag 13 that will be work as output.

We will put data points of lag1, lag11, lag12 in a matrix p

Then we will found scaled matrix of p so that all values lies between 0 and 1, 0 for minimum and 1 for maximum.

Then find the covariance matrix of scaled matrix p

(In this we see how variation takes places among the directions with respect to mean value.

Covariance is such a measure. Covariance is always measured between 2 dimensions.

When we find covariance between a dimension and itself then it called variance.

The above models have different methods of pattern identification and different way of analyzing data, based on their perspective benefits and drawbacks we will choose the one that

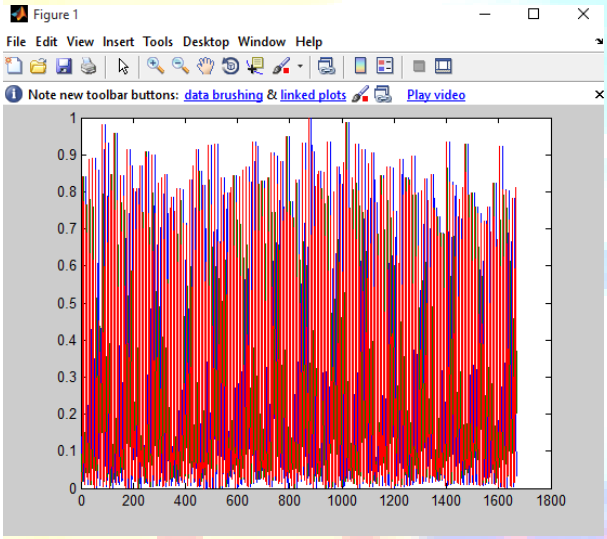
will give us the best result to increase our accuracy while doing prediction of Indian summer monsoon rainfall.)

Now found the eign vector of final matrix. Our principal component of data set will be the vector that has highest eign value.

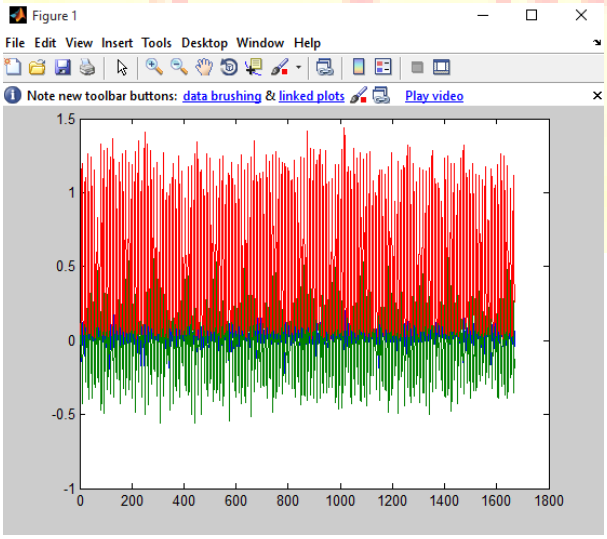
Then we will create new data set or feature vector.

And see how new values giving better result than the past values.

Earlier we found good correlation at lag 1 lag11 and lag 12 so we will see pattern identification before and after applying principle component analysis.



Pattern identification (**Before** applying PCA)



Pattern identification (**After** applying PCA)

We can see clear pattern identification after applying PCA on our prior data

5. Conclusion

In this paper we see that we can found best prediction by using principle component analysis, as it reduces the output dimension and easily found the pattern in data. So among regression model, ANN model and PCA model of prediction, PCA gives the best performance in the rainfall prediction of Indian summer monsoon.

REFERENCES

1-Shraddha Srivastava, , K.C. Tripathi (2012): Artificial Neural Network and Non-Linear Regression:A Comparative Study**,** Internat ional Journal of Scientific and Research Publicat ions.

2-Lindsay I Smith (2012): A tutorial on Principal Components Analysis.

3.K.C. Tripathi ,I M L Das,A K Shai (2006): predictability of sea surface temperature anomalies in Indian ocean using ANN, Indian Journal of Marine Science .

4-Akashdeep Gupta, Anjali Gautam, Chirag Jain, Himanshu Prasad, Neeta Verma(2013) Time Series Analysis of Forecasting Indian Rainfall**,** International Journal of Inventive Engineering and Sciences (IJIES) ISSN: 2319–9598, Volume-1, Issue-6, May 2013

5-P. Guhathakurta1, M. Rajeevan2, and V. Thapliyal2(1999) Long Range Forecasting Indian Summer Monsoon Rainfall by a Hybrid Principal Component Neural Network Model

6. Kumar Abhishek, Abhay Kumar etc,(2012):A Rainfall Prediction Model using Artificial Neural Network, IEEE Control and System Graduate Research Colloquium (ICSGRC 2012)

7-Indira Kadel (2012): Statistical Prediction of Seasonal Rainfall in Nepal**,** APEC Climate Center

8-Kalpesh Patil, M. C. Deo, Subimal Ghosh, andM. Ravichandran(2013):Predicting Sea Surface Temperatures in the North Indian Ocean with Nonlinear Autoregressive Neural Networks. International Journal of Oceanography Volume 2013, Article ID 302479.

9- P. Guhathakurta, M. Rajeevan, and V. Thapliyal(1999):
Long Range Forecasting Indian Summer Monsoon Rainfall
by a Hybrid Principal Component Neural Network Model, Meteorol. Atmos. Phys. 71, 255±266 (1999)

10-**F.** Mekanika, M. A. Imteaza (2013): Capability of Artificial Neural Networks for predicting long-term seasonal rainfalls in east Australia, 20th International Congress on Modelling and Simulation, Adelaide, Australia, 1–6 December 2013