Journal of Data Science 29-50, DOI: 10.6339/JDS.201801_16(1).0002

An Alternative Approach of Handling the Outlier: A Case with Indian NPS

Jitendra Kumar[1], Saurabh Kumar[2], Umme Afifa[3], Ranjan Raj[4]

[1,2,3]*Department of Statistics, Central University of Rajasthan, Kishangarh-305817, Dist.-Ajmer, Rajasthan, India*

[4]*Data Scientist, Nielsen (India) Pvt. Ltd, Bangalore, Karnataka-560068, India*

*Abstract:*Time series modelling is very popular technique used in data science. Main motive of time series modelling is to know the data generating process and also get its parameters which depend on all the observations. There may be few observations which misinterpret the data and also influence the parameters, such type of observations are called Outlier. The present study dealt the handling of outlier in context of ARIMA time series and proposed an alternative approach for the replacement of outlier. In usual process two ways of handling the outlier is popular, in first remove the outliers from the data and second replace it by the nearby values. Removal concept cannot work in the auto-correlated data like time series and similarly replacement of outlier through just previous/after value is also not much appropriate method because of dependency structure. Therefore, we are proposing an alternative approach, in which outlier is replaced by estimated values through best model. Detailed methodology is discussed and then an empirical analysis on the time series of National Pension Scheme (NPS) is carried out. Most of the series are modelled perfectly and few series were not due to non-stationary nature of the series. After getting an outlier free series, forecasting is also done. The realization of the series also performed on proposed methodology to get generalized view of proposed methodology and get similar result.

Key words: NAV, Fund value, NPS, ARIMA

1. Introduction

A time series is a collection of observations generated chronologically. Time series data may be studies in reference of size, dimension as well as parameter of the model. A generating process is always considered as a whole instead of individual numerical field. There may be certain number of observations which affects the model as well as estimator of the parameters. However these observations do not leave permanent impact on the model, such type of observations are called outlier (Tsay, 1986). Main issue with the outlier is to know that particular observation is outlier or not. Tietjen and Moore (1972) discussed the case of outlier detection in univariate data. Hillmer, Bell and Tiao (1983) summarized an iterative outlier detection method. Usually suspected outlier omitted or treated as missing value in normal

course of data analysis and both concepts are not applicable in reference of time series because of inter dependability of observation. Guttman and Tiao (1978), Miller (1980) and Chang (1982) showed that the presence of the outliers may caused serious bias in estimating the Autoregressive (AR) and Moving Average (MA) parameters, when order of the model is known.

A time series analysis done with help of Box Jenkins model (AR, ARIMA and ARMA) and applied to monitor the dynamic severe acute respiratory syndrome (SARS), data recorded daily new cases reported by ministry of health of China (Lai. 2005). The identification of outliers can lead to the discovery of useful knowledge and has a number of practical applications in areas such as credit card fraud detection, athlete performance analysis, voting irregularity analysis, severe weather prediction etc. (Knorr and Ng 1988; Ruts andRousseeuw, 1996; Johnson, Kwok and Ng 1998; Jaykumar and Thomas, 2013). Ljung (1993) detected the additive outlier in ARMA model. Many studies on detecting and modelling in case of outlier, have been extended to the different classes of models including ARIMA, ARCH, GARCH and other models. Some suspected outliers may have large residuals but may not affect the parameter estimates, whereas others may not have large residuals but they affect model specification and parameter estimation. This may lead to misspecified models, biased estimates, and biased forecasts. Thus it is very important in practice to detect outliers and to have a procedure for model building (specification, estimation, and checking) in the presence of outliers.

In the time series modelling, there may be few observations which are affected by several reasons like administrative, political and market changes etc. Due to autocorrelation these fluctuations affect the whole data set which result the poor modelling and biased decision. In the present study we are interested to do the analysis of National Pension Schemes. Pension provision is a way of social security in many countries and covered by public scheme. This is often supplemented by occupational pension schemes. As public scheme varies substantially among advanced economies and working population is limited only around 10 to 25 percent (Schwarz, 2003). In India there are at most 15% working population who was getting pension through old pension scheme and this was not able to facilitate the common people as well as increasing load budget on government expenditure (Gillingham and Kanda, 2001). Majority of public pension schemes are funded by pension assets however others are unfunded like current National Pension Scheme (NPS) is financed from contributions of employees and employers. Occupational pension schemes are implemented through Defined Benefit (DB) and Defined Contribution (DC) schemes. DB schemes provide more benefits after retirement but not permit portability however DC schemes permits portability in case of change of employer. Returns are contributory in respect to accumulated value of retirement fund at the time of retirement but returns under this are closely influence on proposal of

retirement plans, labour markets etc. (Friedberg, 2011). There are several studies on the pension fund industry to discuss the issues, performance, challenges and reforms were required with the old pension fund (Black,1989; Dushi, 2010; Farnzen, 2010; Brown, Clark and Rauh 2011). The concern in this regard is also taken care by many institutions like Department for Works and Pensions report (TSO (Ireland), 2010; Brauninger, 2011), Global Financial Stability Report (March, 2005) and OECD (November, 2005). Main concern of any retirement plan is to become a tool for post-retirement income. Due to contributory nature of the scheme, defiantly ratio of contribution for the scheme under implementation like existing National Pension Scheme (NPS) to provide the expected retirement income. Present NPS scheme is having 10 % contribution of gross income other than perks. Second most important point to be addressed in association of NPS is investment profile, where and how to be invested for getting maximum benefits at the time of retirement. As on today this is not very much clear to the participating employee and also governing institutions are not having transparent mechanism to aware and inform the associate bearers. Recently Sane and Thomas (2013) has addressed well on all associate issues.

Guerrero and Sinha (2004) performed a statistical analysis of market penetration of the Mexican Pension System using generalized logistic curve as well as regression model and concluded that growth pattern of individual pension funds are not similar. Shah (2005) had summarized the Indian pension reforms and goals of pension reforms. As the market is growing and has significant growth on market indices since last two decade but the pension funds are not having growth accordingly therefore there is an urgent need to take necessary action to get similar growth as different saving and investment fund, otherwise this is injustice with those who are forced by the government to contribute their income in NPS fund like state and central government employees of India and also will not serve for common people as ”Swavalamban” and other public welfare pension fund which was targeted from the first day of reform of old pension schemes.

The present study provide alternative approach of handling the outlier in which we are replacing the outlier by most appropriate value from the best model. Usually outliers are excluded from the data, this works only on the concept of linear dependency on independent variable, however in the case of autocorrelation between individual observations this does not work. Therefore, we are proposing the model base replacement. When replacement of outlier is made by its estimated value of best fitted model. An empirical analysis to model the NPS, by NAV and fund value and then proceeded for the further study of outlier identification. For minimizing the impact of the outlier in the model and replaced it by most appropriate values. All outliers are replaced by its estimated value and then forecasted the future NAV and fund values. Study has observed that the conclusion based on the series, free from outlier series is much better than the series contaminated by outlier in the case of NAV

series. However for fund series, we are not able to get outlier free series because of accumulative nature of the observations. The study also reveals that the current NPS is lagging behind 10 % return of saving banks like FD and RD, as RD is similar investment plans.

2. An Approach of Handling the Outlier

An outlier(s) is/are the observations which are not similar with the rest of the observation. Generally two way of handling the outliers are popular in data analysis, first exclusion from the data and second replacement of the outlier by the most appropriate value like nearby values. In case of autocorrelated data like time series data, this is not possible to exclude the observation therefore we use second approach in which outlier is replaced by another value but it is crucial in case of autocorrelated data to get best representation of recorded value. Use of nearby value, may not be perfect representation in case of the process, where dependency exist.

In second case, the nature of the data generating process considering all observations from the same process is more important. However in presence of outlier it is not perfectly realize. In similar way this may affect the model and one may conclude that generating model is different. So handling of outlier must be taken care in the sense of modelling. Therefore, we are using new approach to handle the outlier by replacing outlier with its estimated value. We are using here best ARIMA model for the given data.

A time series y_t follows an ARIMA (p, d, q) process if dth differences of the y_t {t = 1, 2, 3… T} series are an ARMA (p, q) process if we introduced $y_t = (1-B)^d y_t$ then, $\theta_p(B)y_t = \phi_q(B)e_t$. Now substitute for y_t to obtained more closed form for an ARIMA (p,d,q) process as

$$\theta_p(B)(1-B)^d y_t = \phi_q(B)e_t \,;\, e_t \sim N(0, \sigma_e^2)$$

Where, θ_p and ϕ_q are polynomials of order p and q respectively. To estimating the coefficient of the model we are using Maximum Likelihood Estimator (MLE) method. For obtaining the estimated value by using ARIMA model. Flowchart of the analysis is given Figure 1.

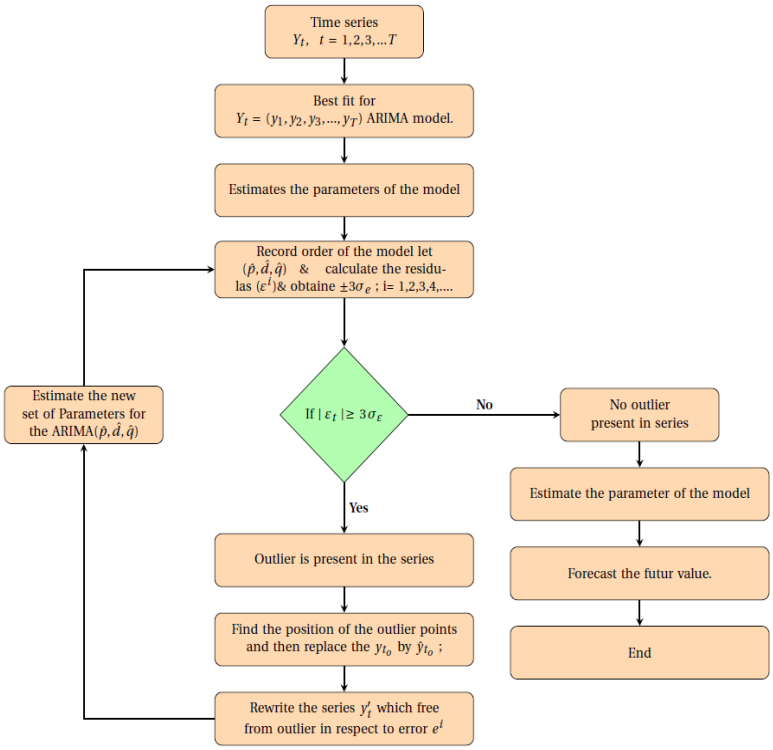

Figure 1: Flowchart of getting outlier free series

As an ARIMA models are autocorrelated model, so replacement of outlier will give a different set of parameters which may again show that there is/are outlier(s) because in respect to new set of parameter some observations may become outlier. Therefore, it is mandatory to continue the analysis till get the outlier free series. Here one very important assumption is made with the data that the specific outlier does not change the process so we have considered the same model. One may explore the case by considering the change of model but we cannot advocate the change of model because individual observation cannot change the process. But if this is case then definitely the process is not stable process and this will create some other modelling complexity in respect to modelling as well as in estimation of parameters.

3. Empirical Analysis of National Pension Scheme

India is showing strength since last three decades as emerging economy and wants to foster sustainable and inclusive growth, fight poverty and reduce inequalities (Planning Commission 11th Plane, 2005). Government also wants to provide a social safety to the

people, they will not fall back into poverty again and if substantial income reductions is there due to any social risk and health risks including old age problem (Krishan, 2010). Pension reforms were taken place in India after the globalization of economy with the objective to start giving more opportunity to all employees in more transparent and contributory manner. NPS was accepted by central government in January 01, 2004 and later on by different state government. This was allowed to all citizens in May 01, 2009 through Swavalamban Yojana. Present NPS is fully dependent on the market investment through different secure and non-secure fund scheme managed by Pension Fund Regulatory and Development Authority (PFRDA) through an operational interface between PFRDA and other NPS intermediaries like Pension Funds, Annuity Service Providers, Trustee Bank etc. It has appointed NSDL e-Governance Infrastructure Limited as Central Record keeping Agency (CRA) for NPS Lite which was established by the Government of India on 23rd *August* 2003 to promote old age income security by establishing new contributory pension plan with the aim to protect the interest of subscribers.

Main motive behind establishing the NPS was to give more opportunity to employee in term of benefit and reduce the financial burden of the government. It was initiated in such a manner which offers a basket of investment choices and Fund managers where subscribers have choices to invest with one or more Central Record keeping Agency (CRA), several Pension Fund Managers (PFMs) and different categories of schemes. PFMs share a common CRA infrastructure and invest money in three asset classes: Asset Class E-Investment in predominantly equity market instrument, Asset Class C- Investment in fixed income instrument other than government securities and Asset Class G-Investment in government securities.

Any Indian citizen between the ages 18 to 55 is eligible to make investment in these schemes. To start investment in NPS, one need to get registered by opening an account at one of the designated Point of Presence (PoP) and then after Permanent Retirement Account Number (PRAN) issued by Central Record keeping Agency (CRA). This is managed through two types of NPS Accounts;

- Tier I - A Non-withdrawable account to which Subscriber shall contribute his/her savings for building a retirement corpus. This is mandatory.
- Tier II - A voluntary savings facility which provides liquidity to subscribers, -i.e. subscribers will be free to withdraw their savings whenever they wish. This is optional. Initial investment for Tier-I is Rs.500/- and Initial investment for Tier-II is Rs.1000/-.

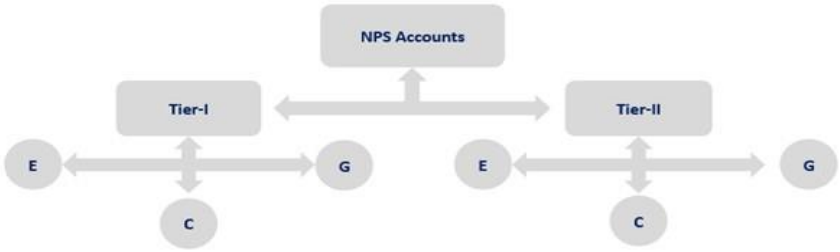

Figure 2: NPS Account Structure

The fund contributed by investor in this account will be invested under the following Asset class:

i. Asset class E (Equity Market Instruments): It is a "High Return - High Risk" fund that invest certain percentage of total fund in index funds i.e. either BSE Sensitive Index or NSE Nifty 50 index.
ii. Asset class G (Government Securities): It is a "Low Return - Low Risk" fund that mainly invest the funds in central or state government bonds or securities.
iii. Asset class C (Credit Risk Bearing Fixed Income Instruments): It is a "Medium Return - Medium Risk" fund that invests in the sector other than stock market and government securities. It mainly invest the funds in public-private sectors such as Public Financial Institutions and Public sector companies, rated municipal bodies/infrastructure bonds.

For notation purpose we follow short Bank name - scheme name - Tier for example ICICI bank scheme E of Tier-l is named by ICICI E1. Present study aims to model the NPS with the proposed methodology for which we consider four different banks namely ICICI, Kotak Mahindra (KM), State Bank of India (SBI) and UTI and have taken the NAV data from 2nd February, 2010 to 30th March, 2014.

Table 1: Summary of NAV values of all four bank schemes (E, C and G)

Series	Mean	Tier-l SD	ADF p – Value	Mean	Tier-ll SD	ADF p - Value
ICICI_E	13.11	1.17	0.682	10.609	0.831	0.726
ICICI_C	13.063	1.805	0.550	12.352	1.662	0.426
ICICI_G	11.874	1.229	0.393	11.133	2.308	0.259
KM_E	12.404	1.034	0.746	11.075	0.891	0.772
KM_C	13.571	1.719	0.635	12.075	1.298	0.622
KM_G	12.153	1.194	0.505	11.192	2.279	0.390
SBI_E	11.335	0.950	0.724	10.539	0.840	0.776
SBI_C	13.557	1.807	0.625	12.177	2.523	0.303
SBI_G	12.958	1.416	0.492	12.177	2.523	0.303
UTI_E	13.071	1.037	0.701	10.400	0.852	0.799
UTI_C	12.292	1.569	0.542	11.943	1.465	0.412
UTI_G	11.827	1.231	0.288	11.598	2.436	0.440

Augmented Dickey-Fuller test is used to check the stationarity of a NAV series and the null hypothesis is taken as series is Non-Stationary. Under this hypothesis we calculate the p- value of each investment scheme for both Tier of all four banks and recorded in Table 1. We observed that p- value of all schemes in both Tier's is greater than 0.05 so we can say that series is Non- Stationary. Here we have also observed that variation in Tier-ll account is more as compare to Tier-l. On the basis of NAV value we have modelled the Time series under the consideration of any abnormal observation present in NAV series.

3.1 Modeling of NAV

Time series observations are recorded chronologically for fix span of time and ARIMA study considers linear dependency of observation on past. We have modelled the series using ARIMA approach and study the error under consideration of outlier. After model fitting, analysis of error is carried-out to exclude the outlier from the series. The rarely happened events, which make sudden ups and downs in the series is called outlier. Time series data is auto correlated data, if outlier is not taken into account, then series managed itself but misguide to the wrong interpretation about the fact related with the model as well as data generating process. Therefore, management of these extreme observations are most important before concluding the analysis as well as performing the other statistical procedures. We have detected the outlier points using $\pm 3\sigma$ limit of the residuals where σ is residual variance. If absolute value of the residual is greater than the value of 3σ then

corresponding observation is taken as outlier in the particular series. These identified outliers have a particular position in the series which affects the process suddenly and due to auto-correlation it cannot be excluded from the data so we have replaced by its estimated value. After replacing the outlier, estimated new set of parameter for the updated data. This may not be outlier free, therefore iterations will be continued till get outlier free series. During the analysis of the Tier-I and Tier-II, we also get some series in which after replacing the outlying observation by its estimates we did not get the outlier free series up to infinite iteration however we have consider maximum number of iteration six hundred, this may be due to wrong model selection or nonlinear model dependency. Suitable model is obtained based on the AIC and reported in Table 2 and Table 3 for Tier-I and Tier-II respectively.

Table 2: Best fitted ARIMA model of Tier- l NAV value Under study

Series	p	d	q	Nout	AIC	# iteration	AIC*
ICICI_E1	0	1	0	1	93.197	2	78.007
ICICI_C1	0	1	1	1	61.394	Inf	**
ICICI_G1	2	1	1	1	14.517	2	-37.198
KM _E1	0	1	0	1	83.673	2	66.650
KM_C1	0	1	1	1	47.617	19	-93.939
KM_G1	0	1	1	1	07.872	2	-37.622
SBI_E1	0	1	0	1	71.057	2	57.932
SBI_C1	0	1	1	1	54.097	Inf	**
SBI_G1	0	1	1	2	19.295	Inf	**
UTI_E1	0	1	0	1	87.913	2	70.891
UTI_C1	0	1	0	1	48.433	218	-421.299
UTI_G1	0	1	1	1	19.074	Inf	**

* After removing outlier from the original series. ** Not getting outlier free series.

We have analyzed under the assumption that model is not changing due to outlier. In our analysis, we found that after certain iterations, series SBI C2 and UTI G2 becomes non stationary for the best order of the original series.

Table 3 : Best fitted ARIMA model of Tier- ll NAV value under study

Series	P	d	q	Nout	AIC	# iteration	AIC*
ICICI_E2	0	1	0	1	065.527	2	53.489
ICICI_C2	0	1	1	1	048.475	2	-56.332
ICICI_G2	0	1	1	2	135.301	Inf	**
KM_E2	0	1	0	1	063.787	2	49.064
KM_C2	1	1	0	1	021.404	35	-125.29
KM_G2	0	2	1	1	133.721	Inf	**
SBI_E2	0	1	0	1	058.071	2	45.085
SBI_C2	0	1	1	1	038.460	****	
SBI_G2	0	2	1	1	140.201	Inf	-181.785
UTI_E2	0	1	0	1	063.351	2	50.085
UTI_C2	0	1	1	1	038.459	15	-90.297
UTI_G2	0	1	1	1	038.460	****	

* After removing outlier from the original series. ** Not getting outlier free series.
*** Non stationary after removing outlier point

We have forecasted the NAV and got significant improvement on reducing the error band and AIC value are listed in the Table 2 and Table 3. Outlier free series are forecasted for both Tier-I and Tier-II which are shown in Figure 3 and Figure 4.

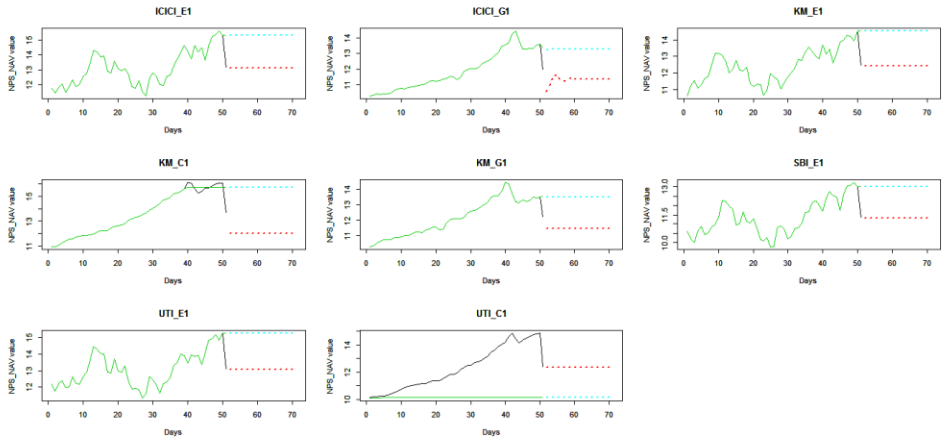

Figure 3: NAV Forecasting of Tier-1, black line stands for series having outlier or original series, green line denotes fitted series, red dotted line says that forecasted value from the original series and blue dotted line represent forecasted after removing outlier.

From Figure 4, one may conclude that if outlier is replace from the series very soon then series have good forecasted value as compare to other series. UTI_C1 series become outlier free series after many iterations and obtained new series which have totally different pattern as compare to the original series. Here, due to autocorrelation at every iteration, got different set of parameters.

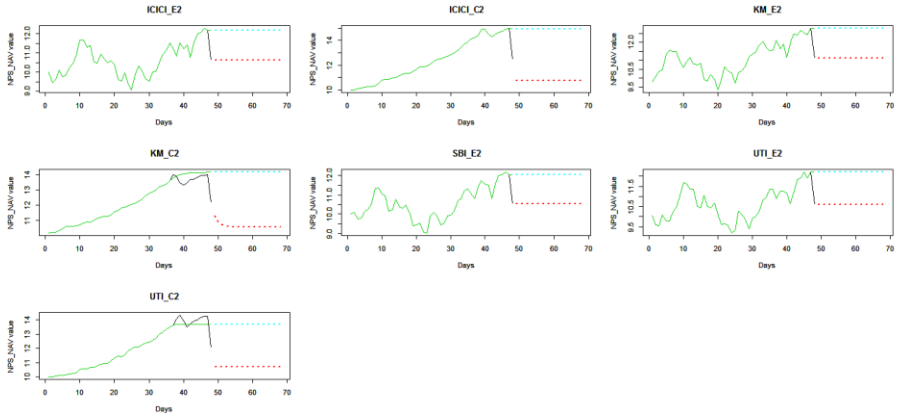

Figure 4: NAV Forecasting of Tier-II, black line stands for series having outlier or original series, green line denotes fitted series, red dotted line says that forecasted value from the original series and blue dotted line represent forecasted after removing outlier.

Therefore series may be completely different and given us a modified series which is completely free from outlier. Main drawback of the process is that at the final stage of the modelling we have a series with a very less original data. In case of more iteration like UTI_C1 where maximum iteration are 218 towards a series of 52 observations. Here one point may be arises that how efficient is this updated outlier free series through model. But this will be better representation of the system because it is originated with the actual data.

3.2 Modelling of Fund Values

In this section, we have considered the fund value for the period under study. Fund value depends on the monthly investment amount invested by the individuals. If any investor invest Rs.1000/- per months in NPS funds. On the basis of that amount we calculate the fund value series, during the calculation we follows the following steps;

1. At first we recorded Average E, G and G, NAV value for all four banks.
2. Fixed the monthly investment amounts Rs.1000 /- (say)

3. To calculate the No. of units = (Investment amount / NAV values)
4. To calculate cumulative number of units in each schemes.
5. Fund value = cumulative units * NAV values
6. Current fund value = current cumulative units * current NAV values.

NAV is aggregate value represents the financial constant in respect of unit participation, therefore we have also modelled the fund value considering monthly investment of Rs.1000/- and get the best model for Tier-1 and Tier-II which are listed on Tables 4 and Table 5 given below;

Table 4 : Best fitted ARIMA model of Tier- l Fund value

Series	p	d	q	Nout	AIC	# iteration	AIC*
ICICI_E1	0	1	0	2	801.665	323	-30.020
ICICI_C1	0	1	1	4	674.176	45	638.367
ICICI_G1	0	1	1	4	642.569	44	575.192
KM E1	0	1	1	2	793.633	67	683.183
KM C1	2	1	2	0	607.914	1	607.914
KM G1	0	1	1	1	640.185	40	567.150
SBI E1	0	1	0	3	795.669	210	-30.020
SBI C1	0	1	1	1	687.756	98	585.439
SBI G1	0	1	1	1	648.020	98	545.703
UTI E1	0	1	0	3	794.136	277	-30.020
UTI C1	0	1	1	5	671.921	59	615.439
UTI_G1	0	1	1	3	641.858	56	578.102

Table 5 : Best fitted ARIMA model of Tier- ll Fund value

Series	p	d	q	Nout	AIC	# iteration	AIC*
ICICI_E2	0	1	0	2	746.82	31	676.859
ICICI_C2	0	1	1	4	630.998	16	603.033
ICICI_G2	0	1	1	2	610.777	Inf	**
KM_E2	0	1	1	1	739.04	8	720.042
KM_C2	0	1	1	4	626.603	68	576.796
KM_G2	0	1	1	1	609.189	Inf	**
SBI_E2	0	1	0	2	743.199	198	-33.495
SBI_C2	0	1	1	3	630.842	29	596.388
SBI_G2	0	1	1	2	615.395	Inf	**
UTI_E2	1	1	0	0	762.187	1	762.187
UTI_C2	2	1	0	1	645.05	***	
UTI_G2	2	1	0	0	645.052	1	645.052

* After removing outlier from the original series. ** Not getting outlier free series.
*** Non stationary after removing outlier point

Considering outlier in observed time series, best model is obtained by the methodology discussed in previous section, which are reported in Table 4 and Table 5 of different pension fund of Tier-I and Tier-II respectively. Based on the best suitable model, we have forecasted fund value for both series: with outlier (Original) as well as without outlier (Outlier Free) series which are shown in Figure 5 and Figure 6 for Tier-I and Tier -II respectively.

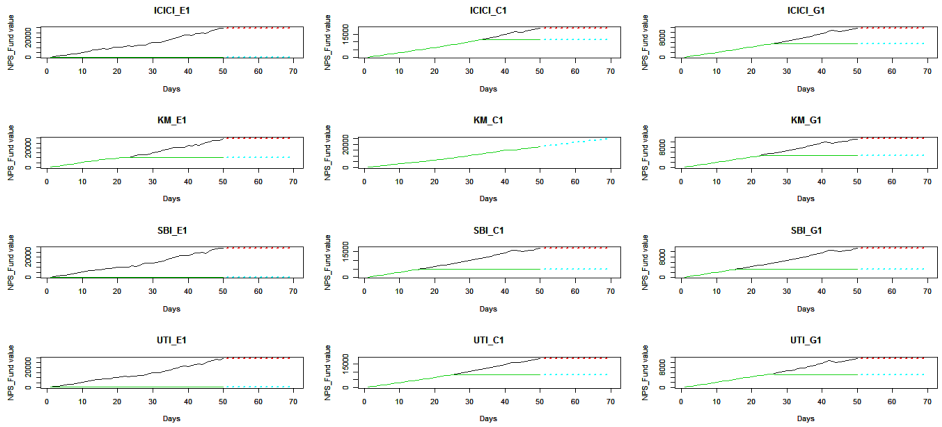

Figure 5: NAV Fund value forecasting of Tier-I, black line stands for series having outlier or original series, green line denotes fitted series, red dotted line says that forecasted value from the original series and blue dotted line represent forecasted after removing outlier.

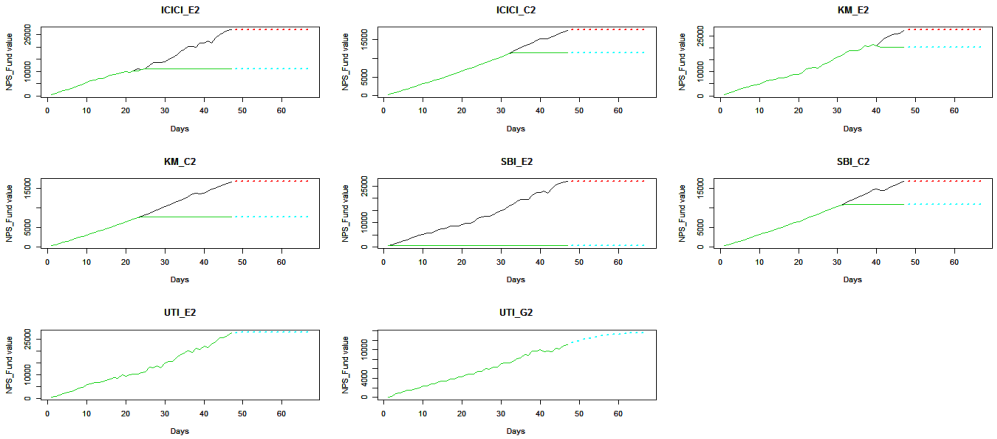

Figure 6: NAV Fund value forecasting of Tier-II, black line stands for series having outlier or original series, green line denotes fitted series, red dotted line says that forecasted value from the original series and blue dotted line represent forecasted after removing outlier.

3.3 Current Value of the NPS

NAV is aggregate value which is a representative value of respective fund. However for common people, value of fund is more important. We have obtained current value of the fund through the NAV series and fund value. In the case of fund value modelling for both Tier- I and II resulting account having lesser amount forecasted value in comparison to NAV Modelling this may be due to wrong model selection, observations may not independent and increasing or decreasing trend may be happened in the series. Any one of the causes may affect the forecast value. Performance of recorded NAV series is poor because it is giving only investment amount and some fund increases maximum around Rs.15,000/-. UTI_G is showing negative growth in the case of NAV value in both tier and remaining series have positive growth.

Table 6: Current Value of National Pension Scheme (NPS)

Series	NAV series				Fund Value series			
	Tier-I		Tier - II		Tier-I		Tier- II	
	With Outlier	Outlier Free	With Outlier	Outlier Free	With Outlier	Outlier Free	With Outlier	Outlier Free
ICICI E	76425.81	84970.65	76287.32	83341.54	29400.67	500	27104.83	11035.59
ICICI C	-	-	70475.98	87134.66	18873.5	11524.43	17518.81	11360.62
ICICI G	74285.62	82622.38	-	-	11495.75	5546.88	-	-
KM E	76487.71	85096.59	76479.1	83716.08	29891.93	10072.53	27508.73	20363.53
KM_C	70703.52	85082.39	70255.30	84874.77	27129.93	27129.93	16726.45	7600.25
KM_G	73692.61	83882.20	-	-	11481.08	4724.11	-	-
SBI_E	76453.98	83882.2	76384.14	83165.25	28874.86	500.00	27017.84	500.00
SBI_C	-	-	-	-	17059.44	4907.2	16763.97	10938.28
SBI_G	-	-	-	-	11372.96	3271.46	-	-
UTI_E	76336.04	84852.8	76346.18	83538.92	29343.59	500.00	27861.91	27861.91
UTI_C	77068.74	67800.84	71651.6	83648.54	18527.19	8203.37	-	-
UTI_G	-	-	-	-	11608.42	5207.7	13781.03	13781.03

In the study of fund value, ARIMA model is not suitable fit for both Tier of fund value series. This may be due to observations that are highly correlated with the past observations due to cumulative nature and proven to get a non-stationary nature of the series. The forecast value of fund series upto 75 months, only few series having similar amount in the case of original series and outlier free series in the Tier-l and Tier- ll series. Most of the series in Tier-l and Tier-ll having fund value less than the original series as compare to the outlier free series. Forecasted fund value after 75 months are recorded in Table 6. In accumulated fund value series, all individual values has cumulative effect which is proven more precisely that the series is non stationary therefore in such case ARIMA may not be best approach for modelling.

4. Realization of Study

Main target of present paper is to develop a methodology to handle the outlier in time series data because exclusion of outlier is not fit in appropriate manner due to autocorrelation. Therefore, such abnormal observations must be integrated in the analysis and proposed methodology is getting the mechanism to analyse the series by replacing the outlier from estimated observation. A simulation study is carried out to get better justification of proposed methodology. Simulation is the realization of real world situation by generating the data under controlled condition. Present section dealt the realization of real data series after the replacement of outlier. We have generated series by using values of coefficient of the time series for five class of model by one type of ARIMA process for the best estimated models, listed in Table 2 to Table 5 in section 3. The details about the steps to be followed in this process are given below:

Step 1: Simulate time series of size T for considered models and taking the seed values from the estimated values of coefficient(s) of best model.
Step 2: Inject an outlier (e_t) at T_0.
Step 3: Fit ARIMA model
Step 4: Obtain the error vector and identify the outlier. (The residual values which is beyond) let it is exist at T_0.
Step 5: Replace by estimated value ( ) and then record new series is

Step 6: Repeat step 3 to 5 till get an outlier free series.
Step 7: Estimate the parameters from the outlier free series.

In simulation study, use R- software (R version 3.3.3; 2017-03-06). The present study have the modelling of monthly NAV of Tier-I and Tier-II. The numerical finding or respective simulation for both type of series are given below in detail.

4.1 NAV Series

In realization of series, only five type of ARIMA models were recorded among the fitted model of all 24 series of Tier-l and Tier-ll account of all four banks. For numerical simplification, we have performed simulation for these five models as given in Table 7. The observed order of model, coefficients values with error variance, number and position of outliers are recorded in Table 6 for respective NAV series.

Table 7:Selected ARIMA model with outlier from NAV series for realization

Model	Order	Bank Name	ar1	ar2	ma1	ma2		#out	T_0
M1	ARIMA(0,1,0)	ICICI_E1	-	-	-	-	0.602	1	51
M2	ARIMA(1,1,0)	KM_C2	0.5599 (0.1186)	-	-	-	0.014	1	39
M3	ARIMA(0,1,1)	UTI_G2	-	-	0.1472 (0.1251)	-	1.434	2	2, 3
M4	ARIMA(2,1,1)	ICICI_G1	0.2427 (0.2741)	-0.4268 (0.2488)	0.7168 (0.1609)	-	0.064	2	24, 51
M5	ARIMA(0,2,1)	KM_G2	-	-	-0.5261 (0.1484)	-	1.102	2	45

Without loss of generality one may assume the outlier as per augment with e_t= (5, 20, 30). Which is injected at T_0 as exist in the real series (please refer last column of Table 7) and then applied steps 1 to 7. A single series cannot show the real world situation so process has been repeated 5,000 times and recorded the average number of irritation to get outlier free series. The average number of irritation, AIC value and error variance with and without outlier observations are recorded in Table 8.

Table 8:Realization results for NAV series.

Model	Magnitude of outlier	# iteration	AIC		AIC*	*
M1	5	14	257.326	1.365	205.980	0.981
	20	27	395.859	3.467	180.616	0.820
	30	27	451.756	5.055	182.608	0.831
M2	5	41	166.243	0.723	-167.547	0.078
	20	42	369.397	2.852	-166.020	0.078
	30	42	429.218	4.273	-167.051	0.078
M3	5	1	269.309	1.438	259.257	1.347
	20	7	391.871	3.312	267.855	1.438
	30	15	446.574	4.793	263.818	1.443
M4	5	1	203.994	0.884	30.693	0.277
	20	1	397.396	3.263	40.167	0.295
	30	2	456.818	4.875	44.529	0.305
M5	5	1	327.287	2.163	318.399	2.046
	20	33	412.941	3.877	259.451	1.376
	30	33	459.697	5.334	260.849	1.394

* Without outlier series

In Table 8, one can see that magnitude of outlier is directly proportional to number of iteration i.e. as increase the magnitude of outlier then number of iteration increases for making the series outlier free. The value of AIC and residuals variance are lesser than after adjusting outlier observation.

4.2 Fund Value Series

In similar manner, same procedure also applied for realization of fund value series. In Section 2.2, there are only five types ARIMA models were fitted on Tier-l and Tier-ll fund value series. Therefore for realization of fund value, selected one series corresponding to five fitted model which are listed in Table 9 with the order of model, coefficients values, error variance, number and position of outliers are recorded. Here it is noted that UTI_E2 fund value series has no outlier.

Table 9: Selected ARIMA model with outlier for fund value

Model	Order	Bank Name	ar1	ar2	ma1	ma2		#out	T_0
M1	ARIMA(0,1,0)	SBI_E2	-	-	-	-	500.218	2	38, 43
M2	ARIMA(1,1,0)	UTI_E2	-0.539 (0.1221)	-	-	-	590.393	0	-
M3	ARIMA(2,1,0)	UTI_G2	-0.398 (0.1410)	0.268 (01457)	-	-	211.348	1	45
M4	ARIMA(0,1,1)	KM_E1	-	-	-0.243 (0.1546)	-	535.992	2	41, 46
M5	ARIMA(2,1,2)	KM_C1	-0.113 (0.1178)	-0.776 (0.1020)	0.696 (0.0701)	0.914 (0.1787)	93.406	1	35

The value of residual variance is quit high as compare to NAV series so error band corresponding this variance is too wide. Hence, magnitude of outlier is considered by larger value. In simulation, we have taken three different value of e_t= (150, 250, and 350) and added as recorded in real series.

Table 10: Realization results for fund value series.

Model	Magnitude of outlier	# iteration	AIC		AIC*	*
	150	24	780.869	46.860	687.557	26.225
M1	250	44	831.834	66.027	637.726	18.334
	350	43	873.257	87.289	639.824	18.497
	150	3	704.454	27.044	630.473	16.402
M3	250	3	757.145	38.606	630.048	16.346
	350	3	797.841	50.828	629.092	16.235
	150	1	739.865	34.081	683.151	23.272
M4	250	1	788.099	47.209	683.300	23.283
	350	1	827.976	61.769	682.390	23.131
	150	1	676.047	21.225	576.235	10.875
M5	250	1	734.493	31.431	574.331	10.742
	350	1	778.474	42.273	574.156	10.730

* Without outlier series

In Table 10, observed that very less iteration required to get a series which is free from outlier observation except M1 model. This may be due to less variability in residuals series.

In present simulation, it is observed that as the outlier appears, the error band increases therefore there is increase on the area of lying the outlier. So it needs more iteration for making the outlier free series. However it is also recorded that outlier free series has less AIC and error variance and both are decreasing with increase the number of iteration.

5. Conclusion

All financial study conclude based on the secondary data. It is important to know the data generating process close to original. However outlier suddenly affect the process and if it is not handled properly will give wrong interpretation in both modelling as well as inferential decision. Proposed mechanism is proposing an new alternative methodology of handling the outlier. Study reveals the importance of mechanism and implemented it with the NPS data. In the modelling of National Pension Scheme, modelling is carried out four banks namely (ICICI, Kotak Mahindra, SBI and UTI) and for both schemes (Tier-I and Tier- II) with new approach of handling the outlier as excluding the outlier from the data is majorly used methodology, which is not applicable in time series data due to autocorrelation. We have proposed the mechanism of replacing the observation model which is based on whole series under the assumption of same model. The accumulated value of fund is showing better growth in NAV value however these were notice outlier free series. Most of the series are contaminated by outlier. The value of AIC is reduced when we apply proposed methodology. The same result found in NPS series and realization of the series. The value of AIC is lesser than in case of Outlier free series it means the new model is more efficient as compare to

older one.As NPS is series of interest of all employees and also the performance of affect broadly to the employee as well as government therefore the analysis may be used for better planning as well proposing the new investments strategy so contributors/investors may be protected in more responsible manner. Proposed mechanism may be also interested for non-linear time series model as well as for the fully machine modelling technologies like machine languages.

Acknowledgment: First author gratefully acknowledge the financial assistant from CSIR, India for financialassistance and SHIATS, Allahabad for carryout the project.

References

[1] *A Sustainable State Pension: when the state pension age will increase to 66*, Department ofWork And Pensions, TSO Ireland, November 2010. http://www.dwp.gov.uk/spa-66-review

[2] *Ageing and pension system reform: Implications for financial markets and economic policies*, Organisation for Economic Co-Operation and Development, November 2005.

[3] Black, F. (1989), Should you use stocks to hedge your pension liability?, *Financial Analysts Journal*, 10–12.

[4] Brauninger, D. (2011), Retirement pensions & sovereign debt in the euro area: *Report on European Integration* EU monitor *83* August 18, 2011.

[5] Brown, J.R., Clark, R. and Rauh, J. (2011), The economics of state and local pensions, *Journal of Pension Economics and Finance,10 (2)*, 161–172.

[5] Chang, I. (1982), Outliers in Time Series, Unpublished Ph. D. dissertation, *Department of Statistics, University of Wisconsin, Madison*.

[6] Dushi, I., Friedberg, L. and Webb, T. (2010), The impact of aggregate mortality risk on defined benefit pension plans, *Journal of Pension Economics & Finance, 9(4)*, 481.

[7] Franzen, D. (2010), Managing investment risk in defined benefit pension funds, *OECD Working Papers on Insurance and Private Pensions*, OECD 38.

[8] Friedberg, L. (2011), Labor market aspects of state and local retirement plans: a review of evidence and a blueprint for future research, *Journal of Pension Economics and finance, 10 (2),* 337–361.

[9] Gillingham, R. and Kanda, D. (2001), Pension reforms in India, *International Monetary Fund,* Working paper WP/01/125.

[10] *Global Financial Stability Report: Risk management and the pension fund industry*, Chapter III, International Monetary Fund, March, 2005.

[11] Guerrero, V. and Sinha, T. (2004), Statistical analysis of market penetration in a mandatory privatized pension market using generalized logistic curves, *Journal of Data Science, 2, 195 – 211.*

[12] Guttmanl, I. and Tiao, G. C. (1978), Effect of correlation on the estimation of a mean in the presence of spurious observations, *Canadian Journal of Statistics,6 (2),* 229–247.

[13] Hillmer, S. C., Bell, W. R. and Tiao,G. C. (1983), Modeling considerations in the seasonal adjustment of economic Time series, *In Applied Time Series Analysis of Economic Data, ed. A. Zellner, Washington DC: U.S. Bureau of the Census* (1983), 74-100.

[14] Jayakumar, G. D. S. andThomas, B. J. (2013), A new procedure of clustering based on multivariate outlier detection, *Journal of Data Science, 11, 69 – 84.*

[15] Johnson, T., Kwok, I. and Ng, R. (1988), Fast computation of 2-dimensional depth contours, *In Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining, AAAI Press* (1998), 224-228.

[16] Knorr, E. and Ng, R. (1998),Algorithms for mining distance- based outliers in large datasets, *In Proc. 24th VLDB Conference* (1998), 24-27.

[17] Krishna, A. (2010), One illness away: how people escape poverty and why they become poor, *Institute of development studies*, IDS room 221, 15:00–16:30.

[18] Lai, D. (2005), Monitoring the SARS epidemic in China: A time series analysis, *Journal of Data Science, 3, 279 – 293.*

[19] Ljung, G. M. (1993), On outlier detection in time series, *Journal of the Royal Statistical Society*, Series B (Methodological), 559–567. http://www.imf.org/external/pubs/ft/wp/2001/wp01125.pdf.

[20] Miller, R. B. (1980), Comments on: Robust estimation on autoregressive models by Martin, *Directions in Time Series, Institute of Mathematical Statistics, Hagwood* (1980), 255–262.

[21] *Planning Commission: Eleventh Five Year Plan*, Govt. of India, India, 2008.http://planningcommission.gov.in/plans/planrel/fiveyr/11th/11v1/11th-vol1.pdf, 23/11/2015.

[22] Ruts, I. and Rousseeuw, P. (1996), Computingdepth contours of bivariate point clouds, *Computational Statistics and Data Analysis, 23*, 153-168.

[23] Sane, R. and Thomas, S. (2013), In search of inclusion: informal sector participation in a voluntary, defined contribution pension system, *Indira Gandhi Institute of Development Research, WP- 2013-022*.

[24] Schwarz, A.M. (2003), Old age security and social pensions, *Social Protection Unit*, World Bank.

[25] Shah, A. (2005), A sustainable & scalable approach in Indian pension reform: New Delhi, *Rajiv Gandhi Institute for Contemporary Studies*.

[26] Tietjen, G. L. and Moore R. H., (1972), Some Grubbs-type statistics for the detection of several outliers, *Technometrics, 14(3)*, 583–597.

[27] Tsay,R. S. (1986), Time series model specification in the presence of outliers, *Journal of the American Statistical Association, 81*, 132–141.

Jitendra Kumar
Department of Statistics, Central University of Rajasthan
Kishangarh-305817, Dist.-Ajmer, Rajasthan, India
Email:vjitendrav@gmail.com

Saurabh Kumar
Department of Statistics, Central University of Rajasthan
Kishangarh-305817, Dist.-Ajmer, Rajasthan, India.
Email: sonysaurabh123@gmail.com

Umme Afifa
Department of Statistics, Central University of Rajasthan
Kishangarh-305817, Dist.-Ajmer, Rajasthan, India
Email: ummeafifa@gmail.com

Ranjan Raj
Data Scientist, Nielsen (India) Pvt. Ltd,Bangalore, Karnataka-560068, India
Email: ranjan.raj08@gmail.com