Visit us - www.researchjournal.co.in ■ DOI : 10.15740/HAS/IRJAES/8.1/8-14

International Research Journal of Agricultural Economics and Statistics

Volume 8 | Issue 1 | March, 2017 | 8-14 ■ e ISSN-2231-6434

Research Paper

Testing statistical models for forecasting malaria cases in India

■ MANISH KR. SHARMA, PAWAN KR. SHARMA, BANTI KUMAR, SUNALI MAHAJAN AND ARCHANA

See end of the paper for authors' affiliations

Correspondence to :
MANISH KR. SHARMA
Sher-e-Kashmir University of Agricultural Sciences and Technology of Jammu, JAMMU (J&K) INDIA
Email : manshstat@gmail.com

Paper History :
Received : *28.07.2016;*
Revised : *02.01.2017;*
Accepted : *12.01.2017*

ABSTRACT : Malaria is still a big problem for a country like India especially with a huge number of slums and poor people having substandard living habits. The present study was conducted on the basis of secondary data available for malaria cases for the period of 1995 to 2011 to find out the trend for number of malaria cases in India and to forecast such cases for future periods. A number of time series models were created from the available data using the SAS software like linear trend, random walk with drift, simple exponential smoothing, log linear and finally the ARIMA models. The most suitable model was found to be the Log linear model with minimum MSE, RMSE and MSPE of 114402.9, 144675.8 and 5.59744, respectively. The forecast for number of malaria cases in India shown a decrease trend from 1122324 cases in the year 2015 to 778868 in the year 2023.

KEY WORDS : Malaria, ARIMA, ACF, PACF, Log linear model, AIC, SBIC

HOW TO CITE THIS PAPER : Sharma, Manish Kr., Sharma, Pawan Kr., Kumar, Banti, Mahajan, Sunali and Archana (2017). Testing statistical models for forecasting malaria cases in India. *Internat. Res. J. Agric. Eco. & Stat.*, **8** (1) : 8-14, **DOI : 10.15740/HAS/IRJAES/8.1/8-14.**

INTRODUCTION :

Malaria is one of the major public health problems of the country. National Vector Borne Disease Control Programme (NVBDCP) reported around 1.5 million cases of malaria in India, of which 40–50 per cent is due to *Plasmodium falciparum*. Malaria can be cured by effective treatment at its early stages. Delay in treatment may lead to serious consequences including death. Prompt and effective treatment is also important for controlling the transmission of malaria.

Malaria is a life-threatening blood disease caused by a parasite that is transmitted to humans by the *Anopheles* mosquito. Out of 400 species of anopheline mosquitoes throughout the world, only 60 species are vectors of malaria. Whereas, in India 45 anopheline species have been reported, out of which 9 species are malarial vectors. In 2001, there were about 1005 deaths in India out of 2.1 million malaria positive cases from the total population of 1.02 billion. Presently India, being malaria in the control phase, but still contributes 66 per cent of the incidence recorded in South-East Asia. However, from 2000 to 2010, the number of cases has showed a decreased by 28 per cent. But in reality actual mortality rate and the malaria incidence are always greater than reported. The reason for under reporting is mainly due to low annual blood smear examination rate (ABER), as National Vector Borne Disease Control Programme (NVBDCP) has set an ABER target of at least 10 per cent (Acharya *et al*., 2013).

Malaria imposes great socio-economic burden on humanity and about 36 per cent of the world population, *i.e.* 2020 million is exposed to the risk of contracting malaria in approximately 90 countries. World Health Organization estimates 300–500 million malaria cases annually worldwide (National Institute of Malaria Research, NIMR). In the south-east Asian Region of WHO, Out of about 1.4 billion people living in 11 countries (land area 8,466,600 km^2, *i.e.* 6 % of global area), 1.2 billion are exposed to the risk of malaria and most of whom live in India (Kondrachine, 1992). However, the south-east Asia contributed only 2.5 million cases to the global burden of malaria. Of this, India alone contributed 76 per cent of the total cases. Even a century after the discovery of malaria transmission through mosquitoes in India by Sir Ronald Ross in 1897, malaria continues to be one of India's leading public health problems. In the 1930's, a treatise written by Sinton (1935) on 'what malaria costs India recorded that the problem of the very existence in many parts of India was in fact the problem of malaria. In those days, it constituted one of the most important causes of economic misfortune, engendering poverty which lowered the physical and intellectual standards of the nation and hampered prosperity and economic progress in every way. There is very limited information on age and gender specific seasonal prevalence of malaria in different paradigms in the country. In the available studies, age and gender classification used is arbitrary (Das *et al.*, 1997; Prakash *et al.*, 1997 and Dutta *et al.,* 1999). The burden is generally higher in males than females in all age groups.

Medical infrastructure along with human resource indicators in India have been improving like allopathic doctors, dental surgeons, AYUSH doctors, nursing personnel and various paramedical health manpower. There is an increase in the availability of Medical staff over the years. With the improvement in all such facilities, malaria cases will have to be decreased and ultimately we can expect this disease to be eliminated from India like the Polio. Keeping this in view, the present study attempts to forecast the malaria cases in India and tries to find out the probable date of its eradication from the country.

MATERIALS AND METHODS :

The present study is based on the secondary data regarding malaria cases in India. The information on time series data for the period of 1995 to 2011 was generated and compiled from published reports on Health Status Indicators in National Health Profile for different years by Central Bureau of Health Intelligence. The present study attempts to find out the trend for number of malaria cases in India and tries to forecast for future periods. A number of time series models were created from the available data using the SAS software like linear trend, log linear trend, random walk with drift, simple exponential smoothing and finally the ARIMA models. The observations were at equally spaced time points.

A linear trend model is a simplest trend model which can be expressed as

$$\mathbf{Y_t = a + b(t) + e_t,}$$

where, 'a' is the y-intercept; when t = 0, b is the slope co-efficient of the time trend, t is time period, Y is the estimated value for time t based on the model and e_t is the random error of the time trend. The problem with this model arises when there is a serial correlation in the data such that R^2 and slope co-efficient may falsely appear to be significant.

Log linear trend model can be expressed as:

$\ln Y_t = a + b(t) + e_t$ or $Y_t = e^{a+b(t)+e(t)}$. Again, like linear trend model, the log linear trend model should also be checked for serial correlation. Random walk model uses the 'first difference' to predict the next change based on the principle that the predicted change can always be added to the current level to yield a predicted level. A random walk model with drift can be expressed as $Y(t) = Y(t\text{-}1) + \beta$, it assumes that from one period to the next, the original time series merely takes a random "step" away from its last recorded position. 'β' represents the constant term and if it is zero, we will have random walk model without drift. Exponential smoothing is simply an adjustment technique which takes the previous period's forecast, and adjusts it up or down based on what actually occurred in that period. It accomplishes this by calculating a weighted average and assigning exponentially decreasing weights as the observation get older.

A non-seasonal series can often be represented as a process whose differences and autoregressive are moving average as prescribed by Box and Jenkins (1976). Differencing a series is likely to be appropriate when there is finite autocorrelation between adjacent observations. Differencing is considered as a filter in forming the series stationary. The general form of ARIMA (p, d, q) model is represented by

$$\mathbf{Y_t = \text{Ø}_1 Y_{t-1} + \text{Ø}_2 Y_{t-2} + \ldots\ldots + O_p Y_{t-p} - {}^{-}_{1} \pounds_{t-1} + {}^{-}_{2} \pounds_{t-2} + \ldots\ldots\ldots + {}^{-}_{q} \pounds_{t-q} + \pounds_t}$$

where, Y_t stands for the value of a stationary time series at time t and e_t's represent random error being independently and normally distributed with zero mean and constant variance for t = 1, 2…..n; d the degree of differencing and O's and θ's are co-efficients to be estimated. The Box-Jenkins method consists of the following steps: Identification, Estimation and diagnostic checking. Identification involves the use of plots of the data, autocorrelations, partial autocorrelations, and other information, to select a simple class of ARIMA models. This amounts to estimating appropriate values for p, d, and q. The parameters of the selected model are estimated using maximum likelihood techniques, backcasting, etc., in the estimation stage. Further, in the third stage *i.e.*, diagnostic checking, the fitted model is checked for inadequacies by considering the autocorrelations of the residual series (the series of residual, or error, values). These steps are applied iteratively until step three does not produce any improvement in the model. ARIMA models thus developed are basically used to forecast the corresponding variable. The entire data is segregated in two parts, one for sample period forecasts and the other for post-sample period forecasts. The former are used to develop confidence in the model and the latter to generate genuine forecasts for use in planning and other purposes.

The study, thus confines to the time series application of SAS software on 12 years previous data on number of malaria cases in India for finding out the time of complete eradication of disease. Ten number of time series models were tried to identify the best model on the basis of root mean square and akaike information criterion (AIC) values. After choosing the best model, it was evaluated for its forecasting accuracy through Autocorrelation function (ACF) and Partial Autocorrelation function (PACF). Thus, the model that has been created was used for forecasting future values.

RESULTS AND DATA ANALYSIS :

In the present study, the data on number of malaria cases reported in India was used for the period 1995 to 2011 to forecast the trend upto the year 2023. The number of malaria cases shows a decreasing trend over time as presented in Fig. 1a of the time series plot. The time series plot clearly indicates that the series was non-stationary which was further confirmed from the Autocorrelation plots as shown in Fig. 1b.

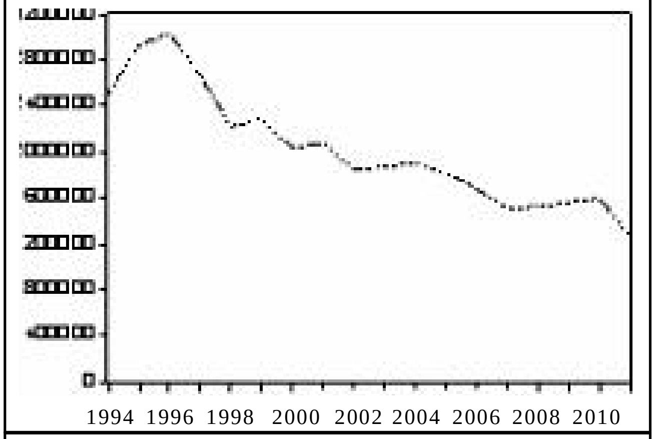
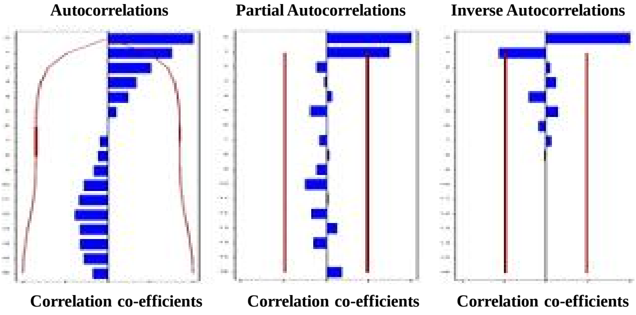


Fig. 1 a : Line graph showing malaria (cases in India)


Autocorrelations

Partial Autocorrelations

Inverse Autocorrelations

Correlation co-efficients

Correlation co-efficients

Correlation co-efficients

Fig. 1 b : Autocorrelation plots

In order to achieve stationary mean, the data was differenced once and the differenced series is presented as Fig. 2a which clearly indicates that the first order difference was enough to make the data mean stationary. The variance however still varies to a small extent. The same can be confirmed through plots of ACF and PACF as presented in Fig. 2b. Therefore ARIMA (p, 1, q) could

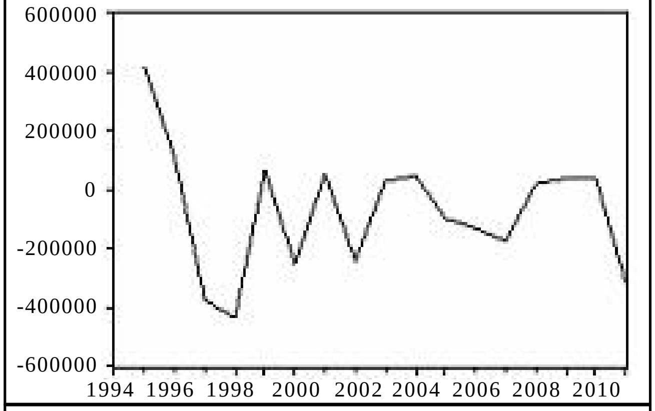


Fig. 2 a : Line graph of first difference of data for malaria cases

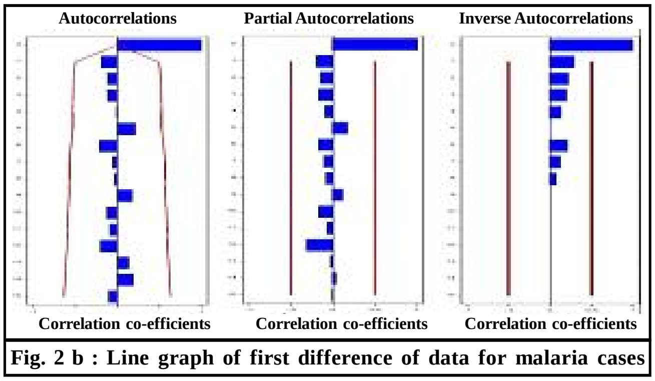


Fig. 2 b : Line graph of first difference of data for malaria cases

be identified for the data at first place.

After identifying the first differenced series of the given data to be stationary, some tentative models including linear trend, random walk with drift, simple exponential smoothing, Log linear and some ARIMA models were considered and built in the software. The results of all the models are presented in Table 1. The best fitted model is accepted on the basis of minimum MSE, RMSE, as both the indicators are found to be more suitable for forecasting. Considering this criterion, ARIMA (0,1,1) was found to be most suitable at the first instance with minimum AIC, minimum SBIC and maximum R^2 of 0.856, but the co-efficients of all the ARIMA models were found to be non-significant because of which we cannot select any of the ARIMA model. Keeping all the factors in mind, the most suitable model was found to be the Log linear model with minimum MSE, RMSE and MSPE of 114402.9, 144675.8 and 5.59744, respectively.

The best fitted equation on the basis of log linear trend model is:

$$\mathbf{Log_e \hat{Y} = 14.88729 - 0.04566t}$$

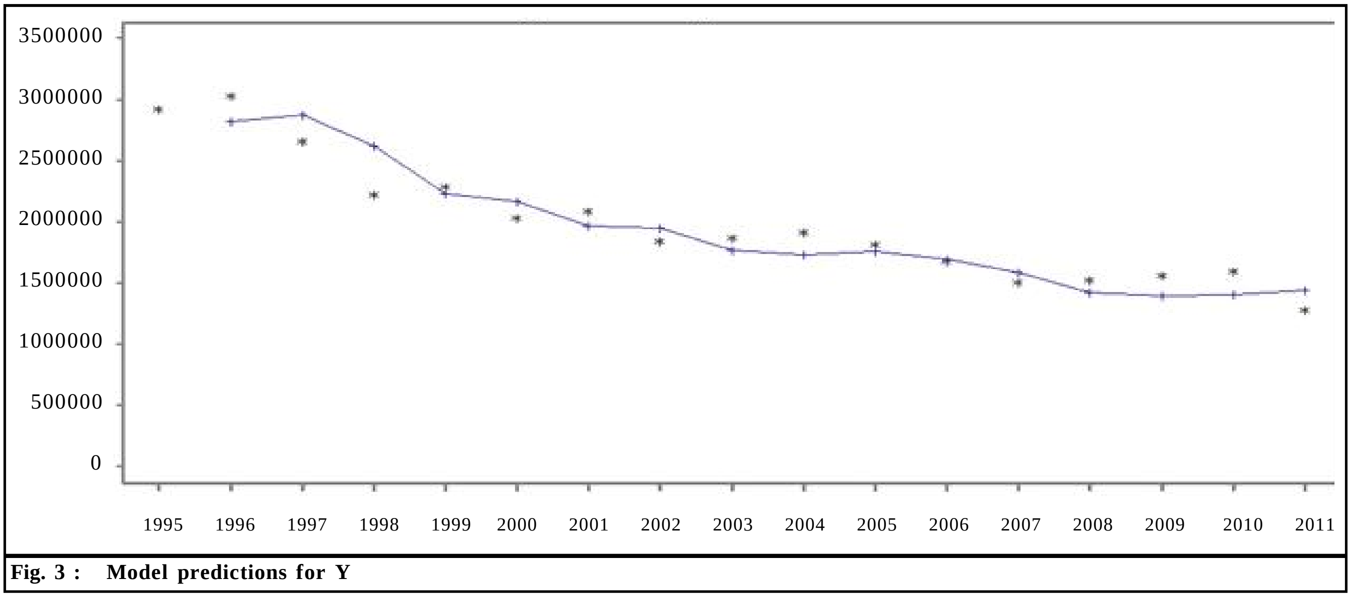


Fig. 3 : Model predictions for Y

Table 1 : Different models generated for forecasting malaria cases in India

Sr. No.	Model	RMSE	MSE	MSPE	AIC	SBIC	Random walk (R^2)	R^2	Adjusted R^2	Prediction criteria
1.	Linear trend	171792.5	134608.1	6.51671	413.83745	415.50387	-	0.875	0.867	3.73827
2.	Random walk with drift	171704.7	152167.8	7.88605	387.71300	388.48559	-	0.848	0.848	2.6512
3.	Simple exponential smoothing	194253.2	142689.5	7.53027	416.01520	416.84842	-	0.841	0.841	4.2451
4.	ARIMA(1,1,0)	168099.5	144697.4	7.54686	389.03397	390.57915	-	0.854	0.843	3.6331
5.	ARIMA(0,1,1)	166906.3	142036.8	7.39921	388.80601	390.35118	+	0.856	0.846	3.5817
6.	ARIMA(1,1,0) no intercept	198900.3	157055.2	8.26200	392.41789	393.19048	-	0.795	0.795	4.4836
7.	ARIMA(0,1,1) no intercept	199173.4	1559202.2	8.20722	392.95702	394.50220	-	0.854	0.843	3.6331
8.	ARIMA(1,1,1) no intercept	190024.2	160512.0	8.34137	392.46179	393.23438	-	0.813	0.800	4.6426
9.	Log linear model	144675.8	114402.9	5.59744	407.99652	409.66295	+	0.912	0.906	.

with R^2 = 0.912, AIC = 407.99652 and SBIC = 409.66295

The selected Log linear models have been fitted and on the basis of estimated parameters and the residuals were determined as shown in Table 2.

Model adequacy:

The adeqacy of the best fitted model has been determined through the examination of the model which is done by two methods:

Test for homoscedasticity:

This can be achieved through the Spearman's rank correlation method. Since the value of the Spearmans rank correlation (0.037) was found to be very low, so we concluded that error terms are distributed with common variance.

Normality of error term:

For 15 degrees of freedom and α=0.05 critical region of t is 2.131. It was verified that all the values of e/s_e are within the range -2.131 to +2.131. Therefore, we could say that our data is 100 per cent normal

The residuals were obtained through the comparison of actual and predicted values of malaria cases in India which can also be viewed by plotting the graph as Fig. 3.

Y	e	e/s_e
2926197	0.0476	0.663091
3035588	0.13	1.810963
2660057	0.0436	0.607369
2222748	-0.0904	-1.25932
2284713	-0.0172	-0.2396
2226360	-0.0889	-1.23842
2126980	-0.0171	-0.23821
2032037	-0.0956	-1.33175
1941332	-0.0352	-0.49035
1854676	0.0348	0.484781
1771888	0.0275	0.383088
1692795	-0.001975	-0.02751
1617233	-0.0667	-0.92916
1545044	-0.009689	-0.13497
1476077	0.0602	0.838615
1410189	0.1288	1.794247
1347241	-0.0496	-0.69095

The forecast and predicted values of malaria cases w.e.f. 2015 to 2023 on the basis of Log linear model has been presented in the Table 3.

The forecast for number of malaria cases in India shown a change (decrease) from 1122324 cases in the year 2015 to 778868 in the year 2023. The estimated

Table 2: Comparison of actual and predicted values of malaria cases for given data

Year	Predicted value for Y	Upper confidence limit	Lower confidence limit	Prediction error	Prediction standard error	Model residual	Model residual standard residual standard error
1995	2926197	3211687	2424030	128799	201051	0.0476	0.0718
1996	3035588	3068325	2315827	363058	192076	0.1300	0.0718
1997	2660057	2931363	2212454	106822	183503	0.0436	0.0718
1998	2222748	2800514	2113696	-216517	175312	-0.0904	0.0718
1999	2284713	2675506	2019346	-45669	167486	-0.0172	0.0718
2000	2226360	2556078	1929207	-194570	160010	-0.0889	0.0718
2001	2126980	2441981	1843092	-41496	152867	-0.0171	0.0718
2002	2032037	2332977	176821	-190018	146044	-0.0956	0.0718
2003	1941332	2228839	1682223	-71929	139525	-0.0352	0.0718
2004	1854676	2129349	1607133	60687	133297	0.0348	0.0718
2005	1771888	2034301	1535394	44681	127347	0.0275	0.0718
2006	1692795	1943495	1466858	-7686	121662	-0.001975	0.0718
2007	1617233	1856742	1401381	-108306	116232	-0.0667	0.0718
2008	1545044	1773862	1338827	-18834	111043	-0.009689	0.0718
2009	1476077	1694681	1279065	87497	106087	0.0602	0.0718
2010	1410189	1619035	1221971	189797	101351	0.1288	0.0718
2011	1347241	1546765	1167425	-68481	96827	-0.0496	0.0718
2012	1287104	1477721	1115314	.	92505	.	.
2013	1229651	1411759	1065529	.	88376	.	.
2014	1174762	1348742	1017967	.	84431	.	.

Table 3: Forecasts of malaria cases in India on the basis of log linear trend model

Year	Predicted value for Y	Upper confidence limit	Lower confidence limit	Prediction standard error
2015	1122324	1288538	972527	80662
2016	1072226	1231020	929116	77062
2017	1024365	1176071	887643	73622
2018	978640	1123574	848021	70335
2019	934956	1073420	810167	67196
2020	893222	1025506	774003	64196
2021	853350	979730	739454	61331
2022	815259	935997	706446	58593
2023	778868	894216	674912	55978

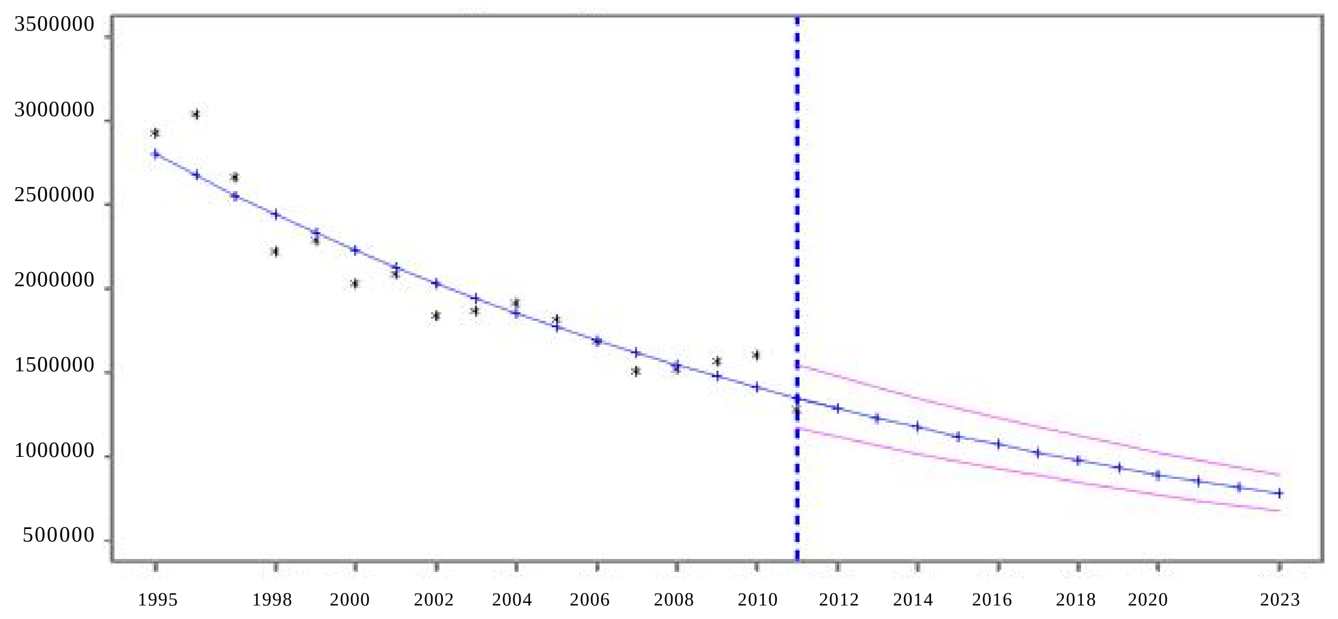


Fig. 4 : Log linear tend plot of forecasts for Y

forecast is also presented in the form of time series plot in Fig. 4.

Conclusion :

This work has been carried out according to the trend of the malaria cases over the years. Out of different models generated for the forecasting of malarial cases in India, log linear model was found to be the best fitted model on the basis of mean squared error criterion, AIC, SBIC and Adjusted R^2. Further, using this log linear trend model, forecasting of number of malarial cases in India has been done from the years 2015 to 2023 which showed a regular decrease in the number of malarial cases in India which is confirmatory with the report "Strategic Action Plan for Malaria Control in India 2007-2012" by Directorate of National Vector Borne Disease Control Programme and Directorate General of Health Services, Min. of Health and FW, Govt. of India.

Authors' affiliations:
PAWAN KR. SHARMA, BANTI KUMAR, SUNALI MAHAJAN AND ARCHANA, Sher-e-Kashmir University of Agricultural Sciences and Technology of Jammu, JAMMU (J&K) INDIA

LITERATURE CITED :

Acharya, A.R., Magisetty, J.L., Chandra, V.R., Chaithra, B.S., Khanum, T. and Vijayan, V.A. (2013). Trend of malaria incidence in the state of Karnataka, India for 2001 to 2011. *Archives Appl. Sci. Res.*, **5**(3):104-111.

Box, G.E. and Jenkins, G.M. (1976). *Time series analysis* *forecasting and control*. Holden- Day. San Fran., USA

Cressie, N. (1988). A graphical procedure for determining non-stationary in time series. *JASA*, **83** (404) : 1108-1015.

Das, N.G., Baruah, I., Kamal, S., Sarkar, P.K., Das, S.C. and Santhanam, K. (1997). An epidemiological and entomological investigation on malaria outbreak at Tamalpur PHC, Assam. *Indian J. Malariol*, **34** (3) : 164–170.

Dutta, P., Khan, A.M. and Mahanta, J. (1999). Problem of malaria in relation to socio-cultural diversity in some ethnic communities of Assam and Arunachal Pradesh. *J. Parasitic Dis.*, **23** : 101–104.

Kondrachine, A.V. (1992). Malaria in WHO Southeast Asia Region. *Indian J. Malariol*, **29** (3) : 129–160.

Makridakis, S. and Hibbon, M. (1979). Accuracy of forecasting: An empirical investigation. *J.Roy.Statist.Soc.A.*, **41**(2): 97-145.

NIMR. Estimation of True Malaria Burden in India. A Profile of National Institute of Malaria Research.

Prakash, A., Mohapatra, P.K., Bhattacharyya, D.R., Doloi, P. and Mahanta, J. (1997). Changing malaria endemicity—a village based study in Sonitpur, Assam. *J. Commun. Dis.*, **29** (2) : 175–178.

Sinton, J.A. (1935). *What malaria costs India*. Malaria Bureau 13. Govt. of India Press Delhi. Health Bull 1935; 26.

WEBLIOGRAPHY

Central Bureau of Health Intelligence (2010). Human Resources in Health Sector. Ministry of Health & Welfare, GOI. http://cbhidghs.nic.in/., accessed on 29th July 2012.

Central Bureau of Health Intelligence (2012). Health Status Indicators, National Health Profile. Ministry of Health & Welfare, GOI. *http://cbhidghs.nic.in*/., accessed on 29th July 2012.