DESIDOC Bulletin of Information Technology, Vol. 27, No. 3, May 2007, pp. 3-12

Exploring Models for the Growth of Literature Data

Anurag Saxena, B.M. Gupta* and Monika Jauhari**

Indira Gandhi National Open University
Maidan Garhi, New Delhi-110 068
E-mail: anurags@ignou.ac.in

**National Institute of Science, Technology and Development Studies*
Pusa Gate, Dr KS Krishnan Marg, New Delhi-110 012
E-mail: bmgupta1@yahoo.com; bmgupta@nistads.res.in

***Bundelkhand University, Jhansi, Uttar Pradesh-284 128*

ABSTRACT

Any time series is forecasted using a suitable model based on the analysis of historical data. The value of a model lies in the efficacy with which it performs the task for which it has been constructed. A model is considered good if it fits the data well. In other words, models should have good parameter values and fit statistics. Many researchers have successfully applied various statistical models to analyse the growth of literature data. However, there is no generalised rule or procedure put forwarded by these researchers. The question, therefore, arises as to how one compares the appropriateness of different type of models fitted to the data? The aim of this paper is to forecast the time series of growth of literature data. Two different approaches to probe this kind of data have been applied. These approaches are the multiplicative seasonal model approach and nonlinear model approach where the trend has an exponential growth form. It has been shown that there is plethora of models that come out with good fit parameters. This communication thus highlights some basic issues related to forecast of growth of literature data.

Keywords: Historical data, time senses, dynamic models

1. INTRODUCTION

A time series is a series of observations taken sequentially over the time. The order in which the observations are included in the data set is very important in a time series. Unlike regression models, it is the ordering property, which distinguishes time series from other series data. Forecasting and time series are also distinct. While a forecast is a view of an uncertain future, the time series is a description of what has actually happened. The aim of this paper is to forecast a time series. This has been

Received 9 February 2007

done with the construction of a suitable model based on the analysis of historical data. Three basic model forms are critical for the description of time series and forecasting. These are trends, seasonal/cyclical variations, and regressions. A dynamic model is a combination of these basic forms with a variance law included. Dynamic models allow changes in parameter values as time passes. The value of information depends heavily on the time it is from. Because of many factors that change with time, any information at a particular point of time would definitely loose its relevance at another point of time. It is equivalent to state that more attention is paid to the more recent data than the older data.

In model building there is a set of rules, but not a hard and fast rule as there is no model, which is the only best model. The value of a model lies in the efficacy with which it performs the task for which it has been constructed. The forecasts made are nothing but probabilistic statements with a condition about the present state of knowledge. Unfortunately, many functions in real-world situations are nonlinear in parameters. Some nonlinear functions can be linearised by transforming the independent and/or dependent variables. But one often encounters functions that cannot be linearised, so the problem of estimating the nonlinear parameter arises. This paper discusses the approaches followed in nonlinear curve fitting and demonstrates a method of linearising and estimating a nonlinear model.

2. DATA: A PRELIMINARY INVESTIGATION

The simplest method to collect the growth data on scientific specialties is either through computerised database on various subject fields or printed abstracting services and bibliographies. There are, however a few databases, abstracting services, and printed bibliographies, which have a long history of development of subject fields. For modelling the growth of data, it is necessary to have a sufficiently large time series. Keeping this in mind, chemical sciences research output as reflected in *Chemical Abstracts* for the period 1901-1994 (referred as CHEM) has been selected.

The series behaves like an exponential growth series. The seasonal component seems to have a random effect. The amplitude increase quite definitely appears to have a direct relationship with underlying series level. The techniques of data transformation is one useful way in which data structures of the amplitude trend relationship kind may be directly addressed. Natural candidates for variance stabilisation in this case are logarithmic transformation and power transformations. The number of publications was at the minimum around 1918. The next low was observed in 1945. From 1918 to 1945, the series behaved like a skewed curve. From 1945 onwards till 1970 it showed an exponential growth. It took a dip around 1972 and then again increase exponentially. Thus, as the level is increasing, the seasonal pattern is varying more prominently. The natural logarithmic series shows an interesting cyclical trend. From 1904 to 1918, it is the first cycle; 1919 to 1945, second cycle; and 1946 to 1993, third cycle. Another prominent thing is the beginning of elevation just after 1993, which could be taken as the beginning of the next cycle.

3. METHODOLOGY USED

Many researchers like Egghe and Rao, Wolfram-Dietmar, Chu-Clara M. and Liu-Xin, Gupta and Karisiddappa[1-3], etc. have applied various statistical series to analyse the data. Few of these models are: Bass model, modified Bass model, Mansfield model, Exponential-logistic model, Gompertz model, and Power model. These models gave good parameter values and fit statistics, but none of the researchers were able to put forward a generalised rule or procedure. West and Harrison[4] described an alternate model for this type of series where the seasonal amplitude increases with level, the multiplicative seasonal model. The model is nonlinear but may be analysed in the linear framework using a suitable transformation. The log-transformed data, in this case, gives a linear trend. Shumway[5] estimated a nonlinear model for the raw earning series with additive trend and seasonal

components, where the trend has an exponential growth form.

$$level_t = \phi level_{t-1} + \omega_t$$

where ω_t is an unknown stochastic term. It is the unknown multiplicative factor ϕ, which make the model nonlinear. The model may be linearised by using an estimate of the growth rate at each time. For forecasting at time t one could set this estimate to be

$$\phi_{t+1} = \{E(level_t/data_t)\}/\{E(level_{t-1}/data_t)\}$$

Operationally, we might simplify this estimate using the online estimated level at time t-1. $\{E(level_{t-1}/data_t)$, in the denominator rather than the one-step back-filtered estimate.

Curve Expert 1.3 has been used for analysing the series. In Curve Expert, the nonlinear models have been divided into families based on their characteristic behaviour. These families and their members are enumerated below.

3.1 Exponential Family

Exponential models have the exponential or logarithmic functions involved. These are generally convex or concave curves, but some models in this group are able to have an inflection point and a maximum or minimum.

Exponential : $y = a*\exp(b*x)$
Modified exponential : $y = a*\exp(b/x)$
Logarithm : $y = a+b*\ln(x)$
Reciprocal logarithm : $y = 1/[a+b*\ln(x)]$
Vapour pressure model : $y = \exp[a+b/x+c*\ln(x)]$

3.2 Power Family

Power family involves raising one or more parameters to the power of the independent variable, or raising the dependent variable to the power of a given parameter. This family is generally a set of convex or concave curves with no inflection points or maxima/ minima.

Power fit : $y= a*x^b$
Modified power : $y = a*b^x$
Shifted power : $y = a*(x-b)^c$
Geometric : $y = a*x^{(b*x)}$
Modified geometric : $y = a*x^{(b/x)}$
Root fit : $y = a^{(1/x)}$
Hoerl model : $y = a*(b^x)*(x^c)$

3.3 Yield-density Models

Yield-density models are widely used, especially in agricultural applications. These models historically have been used to model the relationship between the yield of a crop and the spacing or density of plants. Essentially two types of response are observed in practice: the 'asymptotic' and 'parabolic' yield-density relations. If the response is such that as density (x) increases, but the yield (y) approaches a fixed value, the relationship is asymptotic. If the response is such that there is a distinct optimum as the density increases, the relationship is parabolic. Of course, these types of relationships occur commonly in other scientific areas, therefore, this family of models is very useful.

Reciprocal model : $y = 1/(a + bx)$
Reciprocal quadratic : $y = 1/(a + bx + cx^2)$
Bleasdale model : $y = (a + bx)^{(-1/c)}$
Harris model : $y = 1/(a + bx^c)$

3.4 Growth Family

Growth models are characterised by a monotonic growth from some fixed value to an asymptote. These models are most common in the engineering sciences.

Exponential assoc (2) : $y = a*[1-\exp(-bx)]$
Exponential assoc (3) : $y = a*[b-\exp(-cx)]$
Saturation growth : $y = ax/(b+x)$

3.5 Sigmoidal Family

Processes producing sigmoidal or 'S-shaped' growth curves are common in a wide variety of applications such as biology, engineering, agriculture, and economics. These curves start at a fixed point and increase their growth rate monotonically to reach an inflection point. After this, the growth rate

approaches a final value asymptotically. This family is actually a subset of the Growth Family, but is separated because of their distinctive behaviour.

Gompertz model : $y = a^* \exp [-\exp(b-cx)]$

Logistic model : $y = a/(1+\exp (b-cx))$

Richards model : $y = a/[1+\exp(b-cx)]^{(1/d)}$

MMF model : $y = (ab+cx^d)/(b+x^d)$

Weibull model : $y = a-b^*\exp(-cx^d)$

3.6 Miscellaneous Family

As with many things in life, some things just don't fit into nice categories.

The miscellaneous family is the one in which these 'different' nonlinear regression models live.

Sinusoidal fit : $y = a+b^*\cos(c^*x+d)$

Gaussian model : $y = a^*\exp[-(x-b)2]/(2^*c^2)$

Hyperbolic fit : $y = a+b/x$

Heat-capacity model : $y = a+bx+c/x^2$

Rational function : $y = (a+bx)/(1+cx+dx^2)$

4. ANALYSIS

The analysis of this problem has been classified in three different moulds. All possible models was explored to fit the given series and 5-6 models have been listed, which fits best (in terms of *S* and *r*, where *S* is the standard error of the estimate and *r* is the correlation coefficient). We started by fitting models in the raw series. We fitted five models (Table 1) and were able to explain up to 98 per cent of the variance .

The point, which is worth noticing in Table 2 is that there is a very thin line of difference between these fits and one wonders that why a particular model only is the best model.

The third analysis was done on the basis of Shumway's[6] work where he used the estimate of the growth rate at each time for linearising the model. The analysis of this series used four models and all were able to explain around 99 per cent of the variance.

Table 3 shows that these fits are the most parsimonious among the other discussed earlier. This approach has removed the nonlinearity in the series, and hence giving the best-fit values.

5. DISCUSSION

If one looks this with a critical eye, some of the equations that have been fitted are quite complicated. From this it appears that polynomial models are having some edge over others. To justify the need of transformation and to analyze trends plot of *x* and *y* showing the four best models based on the row series (*Appendix I*), natural log series (*Appendix II*), and Shumway's work (*Appendix III*) were respectively drawn.

Table 1. Fitting models on the raw series

Model	Form	Parameters	*S*	*r*
Sinusoidal fit	$y = a+b\cos(cx+d)$	*a* = 743096.56 *b* = 740091.81 *c* = 0.018 *d* = 0.007	27148.97	0.985
Exponential	$y = ae^{bx}$	*a* = 1.62 e^{-036} *b* = 0.048	38955.413	0.967
Modified power	$y = ab^x$	*a* =1.62 e^{-036} *b* =1.049	38955.907	0.967
Geometric	$y = ax^{bx}$	*a* = 3.76 e^{-032} *b* = 0.0056	39557.278	0.966
MMF (Signoidal model)	$y = \frac{ab+cx^d}{b+x^d}$	*a* = -27634897 *b* = 176.87 *c* = 42069378 *d* = 0.629	67035.88	0.902

Table 2. Fitting models on the natural log series

Model	Form	Parameters	*S*	*r*
Sinusoidal fit	$y = a+b \cos(cx+d)$	$a = 11.407$ $b = 1.996$, $c = 0.029$ $d = -20.991$	0.254	0.981
4^{th} degree poly	$y = a+bx+cx^2+dx^3+ex^4$	$a = 157767.02$ $b = 351.744$, $c = -0.2921$ $d = 0.0001$ $e = -1.4677^{-008}$	0.255	0.981
Power fit	$y = ax^b$	$a = 1.067\ e^{-027}$, $b = 8.516$	0.270	0.977
Modified power	$y = ab^x$	$a = 0.00224$ $b = 1.0043$	0.270	0.977
Exponential	$y = ae^{bx}$	$a = 0.00224$ $b = 0.0043$	0.270	0.977
Geometric	$y = ax^{bx}$	$a = 0.006$ $b = 0.00051$	0.270	0.977

Table 3. Fitting models on the basis of Shumway's work

Model	Form	Parameters	*S*	*r*
3^{rd} degree poly	$y = a+bx+cx^2+dx^3$	$a = 7124.515$ $b = 0.64142$ $c = 1.91\ e^{-006}$ $d = -2.975$	11577.59	0.997
Quadratic fit	$y = a+bx+cx^2$	$a = -3487.44$ $b = 1.039$ $c = -1.88\ e^{-007}$	13889.53	0.996
Linear fit	$y = a+bx$	$a = -256.255$ $b = 0.959$	14143.35	0.995
MMF	$y = \frac{ab+cx^d}{b+x^d}$	$a = -31232.955$ $b = 293244.24$ $c = 5712255.3$ $d = 0.77622423$	18860.379	0.992

Shumway's model, that has been fitted is on the first differences. Thus, we reduced one degree of polynomial fitting. The best approach was to try to plot (after suitable transformation) the first, second and third differences and choose the proper degree when differenced data do not show any trend, (should not be overdone). Usually one or two differences do the job. Once we have got rid of trend then we should try to fit autoregressive or moving average model, which take into account the autocorrelations present in the data. This is called Box-Jenkins's approach. The fitting done here looks quite good but it is usually seen that this approach, is not good for forecasting, whereas in the Box-Jenkins's approach forecast depends more on the recent data than on the past date, and thus giving a better forecast.

6. CONCLUSION

The big question, which automatically arises, is how to compare the appropriateness for the different type of functions fitted to this data. Should one just fit all the commonly used functions and see which one fits the data 'the best'. A good analysis requires robust techniques in assessing and empirically developing the model. The data is never wrong and thus 'statistics' should 'speak' for the data. One should not lead by assumption but should try empirical evidence. The data

itself suggests as to how and in what form the history of the series is to be used. In summary, we have not adhered to pre-specifying the model but tried to develop the model by keeping it simple and parsimonious. Nobody can claim that a particular model is the true equation of the data in question as the true equation is only known to GOD and hence rightly said "All models are wrong, some are useful!"[6].

ACKNOWLEDGEMENT

The authors are thankful to Mr Daniel Hyams, 112 B Crossgate Street, Starkville, MS 39759 for the software Curve Expert 1.3, which has extensively been used for fitting and defining the models.

REFERENCES

1. Egghe, L. & Ravichandra Rao, I.K. Classification of growth models based on growth rates and its applications. *Scientometrics*, 1992, **25**, 5-46.

2. Wolfram, Dietmar; Chu-Clara, M & Liu-Xin. Growth of knowledge: Bibliometric analysis using online databases data. *In* Informetrics, edited by L.Egghe and R.Rousseau. 89/90 Amsterdam, Elsevier, 1990. pp 355-72.

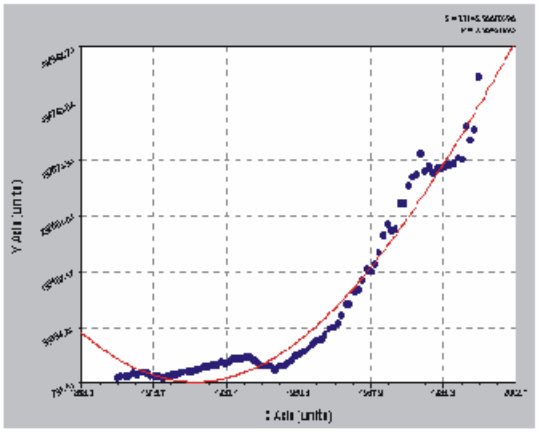
3. Gupta, B.M. & Karisiddappa, C.R. Modelling the growth of literature in the area of theoretical population genetics. *Scientometrics*, 2000, **49**, 321-55.

4. West, M. & Harrison, P.J. Bayesian Forecasting and Dynamic models. Springer-Verlag, New York, 1989.

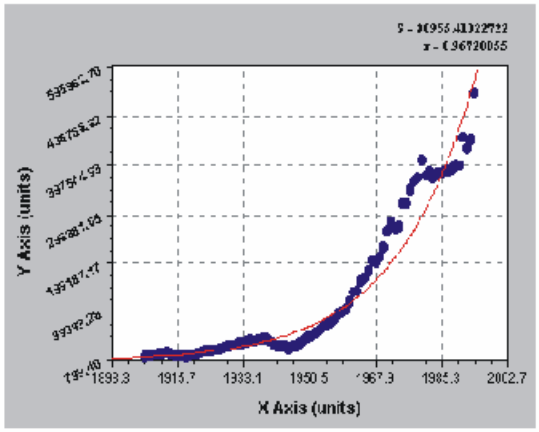
5. Shumway, R.H. Applied Statistical time series analysis, Prentice Hall, Englewood Cliffs, NJ, 1988.

6. Reilly, David P. Newsgroups: Sci.bio. ecology, Sci.Stat.consult., 1997.

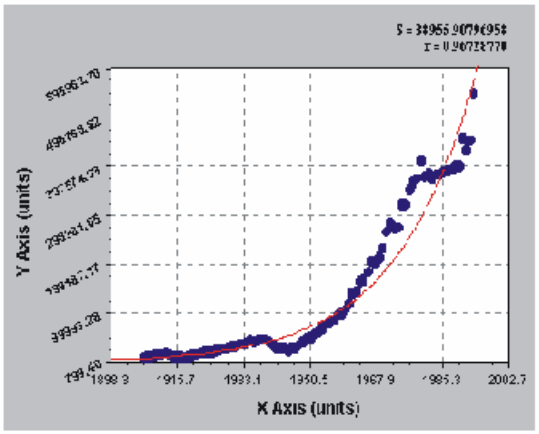
The four best fits based on the raw series



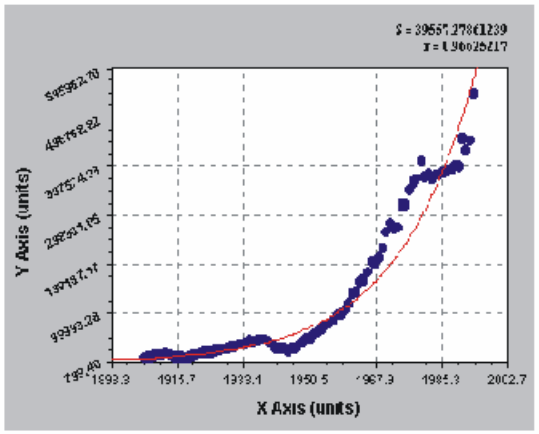
Sinusoidal fit



Exponential

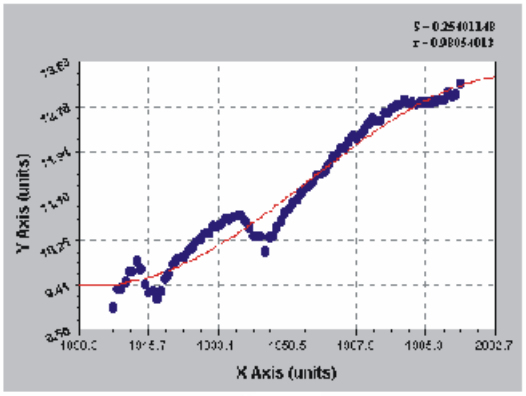


Modified power

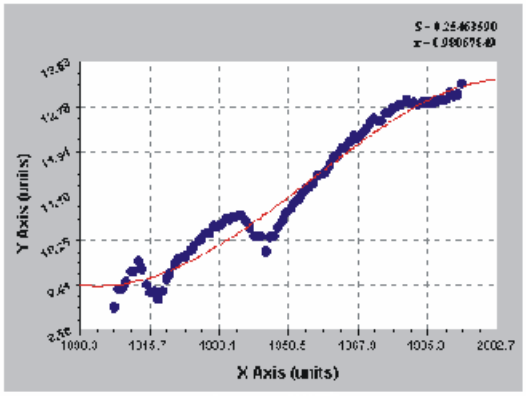


Geometric model

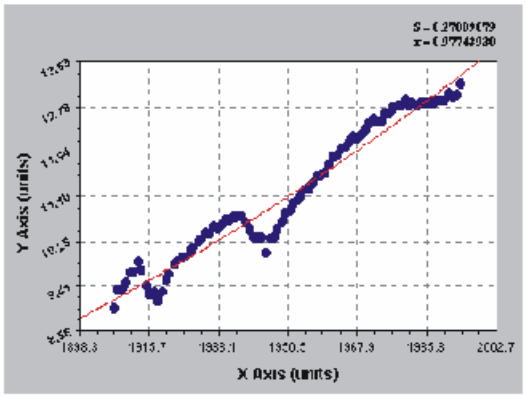
The four best fits based on the natural log series



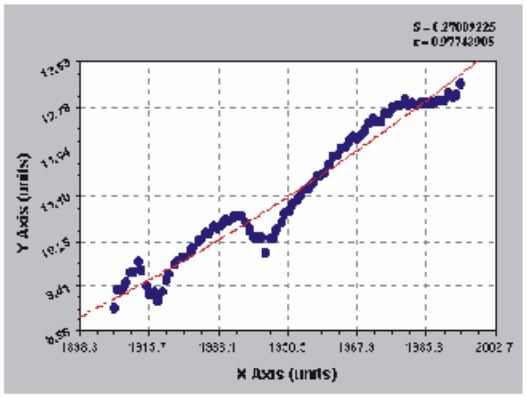
Sinusoidal fit



4th degree polynomial fit

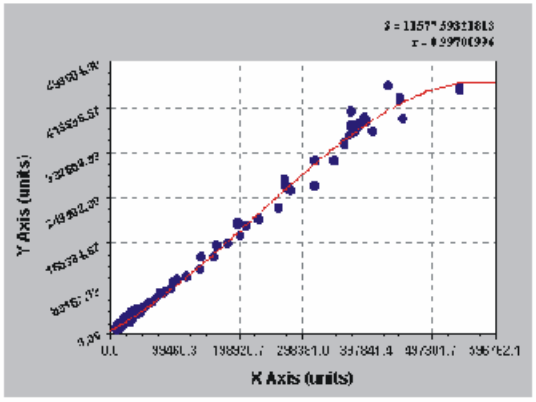


Power fit

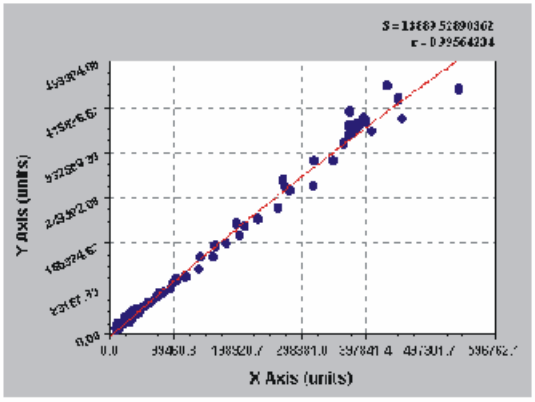


Modified power

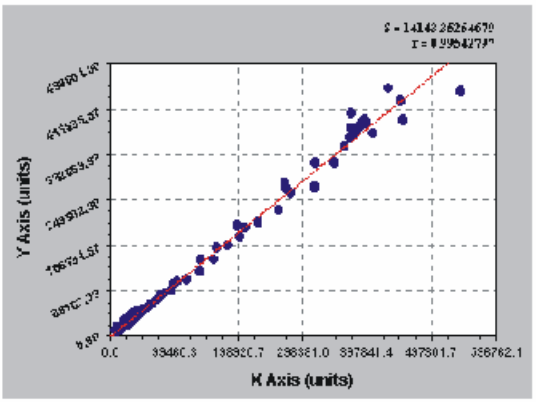
Showing the four best fits based on the series obtained on the basis of Shumway's work



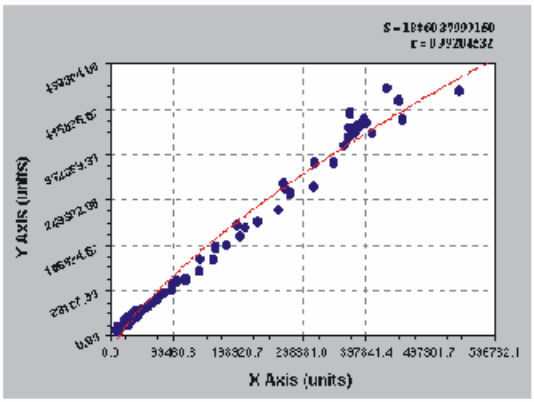
3rd degree polynomial fit



Quadratic fit



Linear fit



MMF Model

Contributors

Dr Anurag Saxena is a Reader in the School of Management at Indira Gandhi National Open University in New Delhi. He has a doctorate degree in Statistics and a postgraduate diploma in Distance Education. He has about 18 Years of research and teaching experience including 14 years in open and distance learning. Current teaching assignments of Dr Saxena include quantitative techniques, information systems for managers, and supply chain management. His areas of interest are distance education, informetrics, webometrics, data-mining, and knowledge management.

Dr B.M. Gupta received his PhD in Scientometrics from Karnataka University in 1999 and is working in this area since last 30 years. His areas of interest include measurement of Indian S&T, international collaboration, productivity of scientists, growth and obsolescence. Presently, he is working as Senior Scientist at the National Institute of Science, Technology and Development Studies, a national institute under Council of Scientific and Industrial Research.

Ms Monika Jauhari is a research scholar at Bundelkhand University, Jhansi. She is a postgraduate in Mathematics and has also done Associateship in Information Science from NISCAIR, New Delhi. Her areas of research include bibliometrics and information science.