Complexity Measure in time-series data: Application to Commodity Export and Research Publication Data

Ashok Jain[1] and M.K. Das[2]

ABSTRACT

Time series data on production, export and research are extensively used in economics. Recognising that time series data is an output of a complex 'system' involving dynamic interaction between several internal and external factors, in physical, biological, and medical sciences the technique of permutation entropy has emerged as a technique to estimate complexity (complexity index) embedded in the system under study. As export and research publication data also represents output of complex systems, this paper measurescomplexity indices of systems associated with a) export of low and high technology products from India and b) research papers published in journals. The analysis shows differences in the complexity indices of systems associated with manufactured

[1] Vice President (Research and Academic Development), EMPI Business School, New Delhi
[2] Institute of Informatics & Communication, University of Delhi South Campus, New Delhi 110021

commodities deployinglow and high technologies (batch and mass production) and single and multiple authored papers in various journals. Permutation entropy technique being independent ofprior model or hypothesis, the results obtained from its application are expected to provide additional perceptive to analysis of high and low technology exports as also to strategic interventions and theory-based modelling work.

1.Introduction

Human societies and economies are complex systems. Time series data is one of the manifestations of the working of such complex systems and facilitates study of systems whether created by humans or by nature. Conventionally time series data is used in conjunction with some theory -based model of the system under consideration.

Complex systems can be broadly divided into three classes [1].

(1) Systems designed to work on the basis causal laws. Complexity in machines is a prime example of such complexity in human-made systems. It is also known from Second law of thermodynamics that entropy in such systems will invariably increase over time and unless periodically 'repaired' or regulated, the system would at some time cease to perform. Control of such systems is possible to a large extent as the internal components and interactions between them are based on known laws or models and thus are often termed as deterministic.

(2) At the other extreme are the inherently Chaotic Systems. The relationship between components of such systems is such that some small local change can get transmitted to a large change in system(sensitivity to initial conditions) and

the causal relationship between what transmits local change to the large change i.e., instability in the system isnot known.

(3) Self-adaptive systems are yet another class of complex systems that is neither totally engineered nor completely chaotic. Such systems increase their survival capabilities over time with some underlying logic that may or may not fit into known rules/norms or models. Nature provides a good example of such systems in which species evolve contrary to second law of thermodynamics, overcoming competition and building on small changes.

(For a rigorous classification of complex systems, one may refer to standard literature on non-linear dynamics).

While time series data is extensively used in conjunction with a theory-based model, little attention is given to the nature of complexity inherent in the system either in model building or in drawing policy conclusions. Symbolic Time Series Analysis (STSA) in data mining has emerged as an approach to provide some understanding of the nature of complexity in social, economic and physical systems. STSA provide complexity measure in terms of symbolic patterns or indices resulting from fluctuations that exist in the data. In other words, rather than looking for correlations amongthe variables the model prescribes, the STSA approach results in a measure of short and long term correlation of patterns in the time series data of a single variable which inherently incorporates theeffect of other variables of the model/system.

For example in economics,Juan Gabriel Brida has applied STSA for a comparative analysis of the national-regional dynamics of accumulation, technical change and employment in Italian economy [2]. Yamamoto et al[3] have also shown the suitability of STSAin data processing in economics for partitioning of resources, competence, information access, and knowledge representation in an integrated methodological design. Earlier Daw et al [4] had extensively discussed the significance of STSA approach in physical system to extract information about the processes that manifest as data out put. In this paper, following STSA approach, we calculate Permutation Entropy from time series as an index of complexity.

The paper is organized as follows: In the next section 2 the mathematical formulation of methodology of analysing data is presented. In section 3 results pertaining to export of different commodities are presented and discussed using data for the period (1987-2013). In section 4, we present results of a similar exercisein non-economic domain of publication data for the period 1950-2005 in the area of astronomy and astrophysics. Conclusions are given in section 5.

2. Mathematical Formulation

In most problems in physics and economics, weare quite often interested in hypothesis testing on the basis of temporal patterns in time series data. If the time series data involves sinusoidalperiodicities, one uses the Fourier transforms to decipher such patterns. However more complex dynamics / behaviour in a time series data such as bifurcation, intermittency and chaoticoscillations, may require more mathematically involved approaches. In this work, we attempt to analyze the time series data of export commodities during 1987-2013 and alsopublication data for the period 1950-2005 in journals of astronomy/astrophysics.

While studying evolution of a dynamical system in a given time domain generally one attempts to find a quantitative measure of information contained the data. If the observable is, $X(t)$, then one of the measures of such information is obtained via the probability distribution $P$of $X(t)$. Assuming that the observable assumes only discrete values; we may associate a discrete probability distribution with M degree of freedom as:

$$P = \{p_i: i = 1,2, \ldots \ldots, M\}$$

Therefore the information content in the observable,$X(t)$, maybe characterised in terms of Shannon'*sentropy* as:

$$S[P] = -\sum_{i=1}^{M} p_i \, ln(p_i).$$

If $[P] = S_{min} = 0$, then it is possible to predict which of the possible outcomes i , with the corresponding probability p_i will actually occur. Therefore, the system is treated as *deterministic*.

To get*minimal information*, the probability function is uniformly distributed among various degree of freedom i.e.,$(P = \{\frac{1}{M}, \forall i = 1, \ldots, M\}) as$:

$$S[P] = S_{max} = \ln M.$$

Following points are to be noted in the above approach:

1. The approach does not take into account temporal relationships that may exist between the values at different points of time. In other words any temporal patterns inherent in the data are ignored and to that extent information about the system that could be derived from the data is lost. For instance, consider two events in time in which the value of the variable goes from 0 to 1.The state 0 or 1 can be reached in various ways i.e.,

$X_1 = \{0,0,1,1,1\}$ and $X_2 = \{1,0,1,0,1\}$

It is readily observed that

$$S[P(X_1)] = S[P(X_2)].$$

2.Further we observe that the classical entropy measure assumes prior knowledge of the system. This implies a probability distribution to be assigned to a time series of the system beforehand. Most of the methods used to find P do not consider the dynamical properties of the system and so are ad-hoc in nature. Therefore Shannon's entropy alone is not sufficient to provide information regarding the dynamical changes taking place in the system.

In recent years symbolic treatment of time series data has been carried out as it is related to the discipline of symbolic dynamics [4]. In fact a raw time series measurements is linked to a series of discredited symbols that are subsequently processed to extract information about the generating process that operate within the system under study. .

Bandt and Pompe [5] suggested a method to find the probability distribution function P, which incorporates the time causality by comparing the neighbouring values in the time series. In the absence of knowledge of the processes that operate within the system (causal information) , this method allows generation of a symbol sequence from a time series that serves as a distinct marker of the processes. In other words, encoding the time series into a sequence of symbols incorporates the causal information. In this method, consider the time series

$$X = \{x_t : t = 1,2 \ldots., N\}$$

At each time $\mathrm{t} = q$, a vector composed of the n-th subsequent values is constructed:

$$q \vdash \left(x_q, x_{q+1}, \ldots., x_{q+(n-2)}, x_{q+(n-1)}\right)$$

Where$n$ is termed as embedding dimension and corresponds to quantum of information contained in each vector. An ordinal pattern defined as the permutation

$$\pi = (r_0 r_1 \ldots. r_{n-1})$$

Of$(01 \ldots \ldots [n-1])$is associated to this vector,such that

$$x_{q+r_0} \leq x_{q+r_1} \leq \cdots \leq x_{q+r_{n-2}} \leq x_{q+r_{n-1}}$$

In other words, the values of each vector are sorted in an ascending order, and a permutation pattern π is created with the offset of the permuted values. For instance, if $x = (4\ , 6\ , 9\ , 10,\ 7,\ 11\ , 3)$, then it is readily observed that for $n = 3$, 2 patterns (0,1,2) correspond to $x_t < x_{t+1} < x_{t+2}$; 2 patterns (2,0,1) correspond to $x_{t+2} < x_t < x_{t+1}$ and 1 pattern (1,0,2) corresponds to $x_{t+1} < x_t < x_{t+2}$resulting in permutation entropy $P_E = 1.5219$.

To illustrate the above approach of linking symbolic dynamics to the time series data to extract information of the nature of complexity inherent in the system under study, we computed the permutation entropy for the following:

(a)Export of commodities for the period 1987-2013.

(b)Research Publication data from 1950-2005 in Astronomy and Astrophysics.

3. Applicationto Indian commodity export data for the period 1987-2013

Data is divided into following categories. Clearly data is output of complex processes that had operated within a system associated with the category.

1. Primary products
2. Manufactured Goods
3. Petroleum Products
4. Others (All commodities)

By simple averaging the category data over 1987-88 to 1996-97 (phase- I), 1997-98 to 2006-07 (Phase-II) and 2007-08 to 20012-13(Phase –III), first cut information on dynamics of exports is readily derived.Figs 1 (a) – (c) show changes in percentage contribution of different commodities to exports.

Following inferences can be drawn:

Share of primary products, manufactured goods, and petroleum products and 'all other commodities' in total commodity exports during phase-I is 74%, 23%, 2% and 1% respectively. The corresponding percentages are   73%, 17%, 8% and 2% in phase-II and 63%, 15%, 8% and 5% in phase III.

Share of primary product in total exports during the period 1987-2013 declined by around 11% andby 10% during the phase-IIIcompared to phase-II. However the export of petroleum product has increased considerably from 2% in phase-I to 18% in phase-III. It may be also noted that the export of manufactured goods declined by around 8% in phase-III with respect to phase-I.

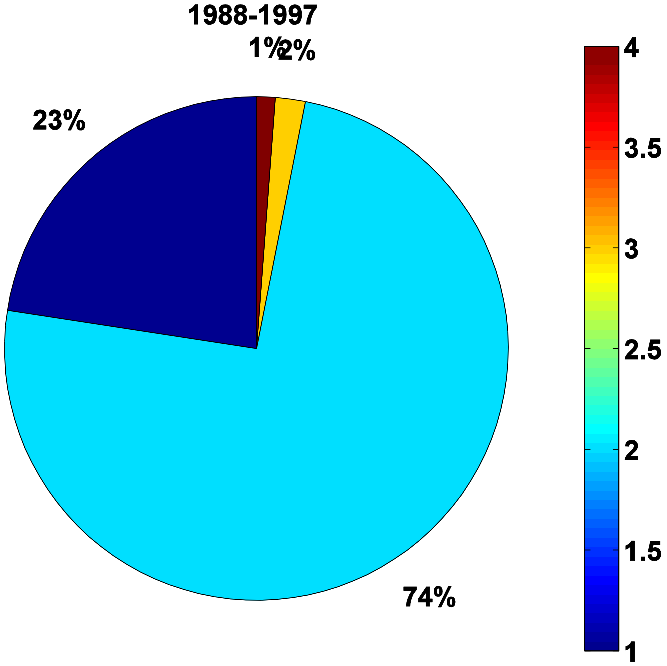

Fig.1 (a), Pie diagram of category data in phase-I

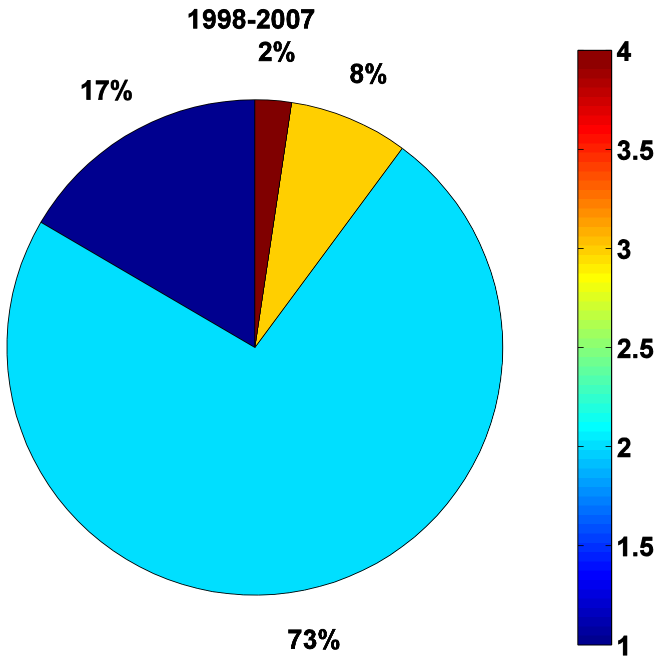

Fig.1 (b), Pie diagram of category data in phase-II

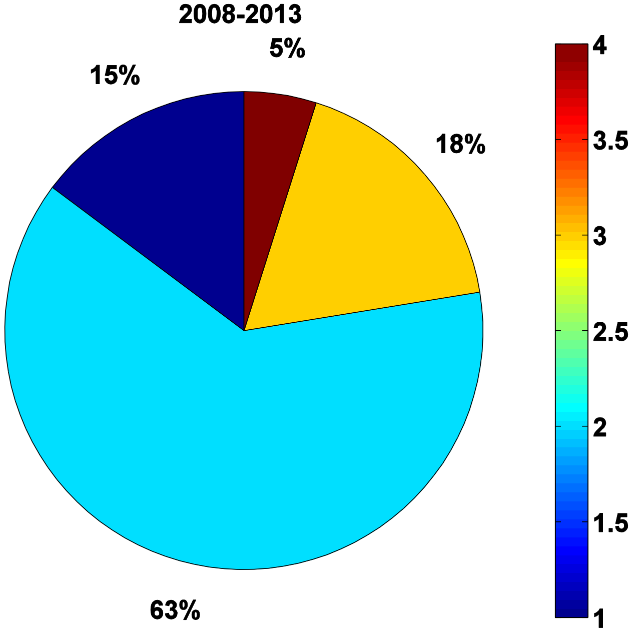

Fig.1(c), Pie diagram of category data in phase-III

Turning attention to full data set, Figure 2 shows actual time series data further classified in category of

- Primary products: Agriculture and allied products(IA) and ores and minerals (IB)
- Manufactured products:Leather and manufactured goods (II-A), Chemicals and related products (II-B), Engineering goods (II-C), Textile and textile products (II-D), Gems and jewellery (II-E), Handicrafts (excluding hand made carpets (II-F) and ‘other manufactured goods’ (II-G)
- Petroleum products (III) and
- Others (IV).

Although most data sets, except for few cases, show an increasing trend, fluctuations in data are different in different categories. By carrying out complexity analysis described earlier in section: 2 we shall show that the fluctuationsindicate differences in the complexity in processes inherent in the systems associated with export of different items.

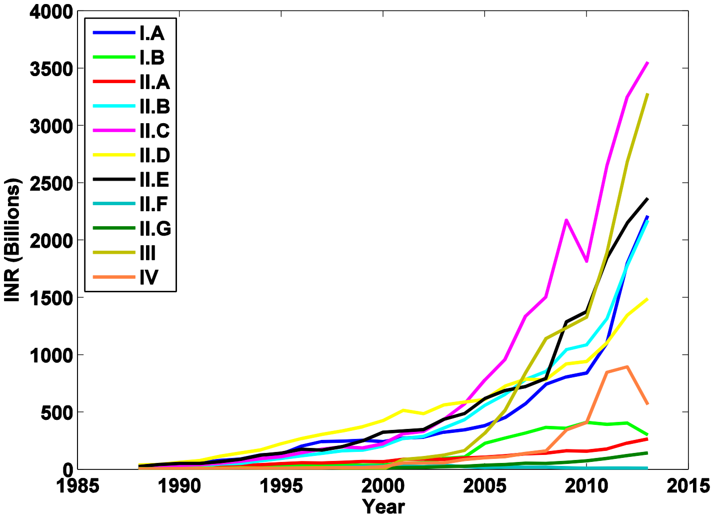

Fig.2 Time series for various export commodities

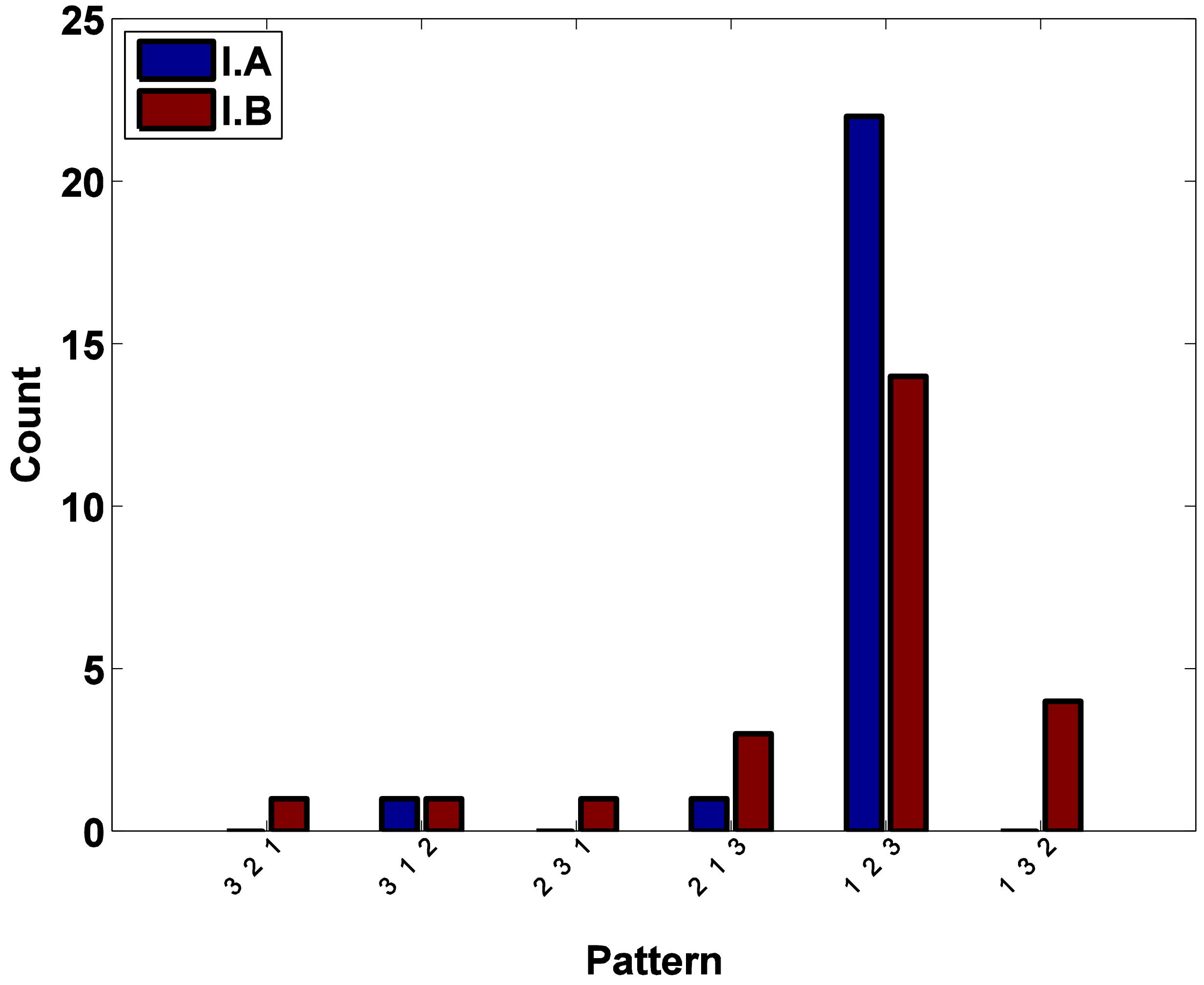

Fig.3a Pattern counts in various time series of Fig.2.

Taking $n = 3$, (two neighbouring values of variable) the various possible symbolic patterns in a time series could be (1,2,3), (2,3,1), (3,1,2), (1,3,2), (3, 2, 1) and (2,1,3). The results of estimated pattern statistics are provided in Fig.3(a)-(d).

It is readily observed from the Figures that the occurrence of pattern (1,2,3) dominates all the time series (increasing trend). However several other patterns also occur in time series of different items.

The complexity in time series data of different items, as expressed in terms of permutation entropy, P_E,[6,7] has been computed and results are shown in Table:1

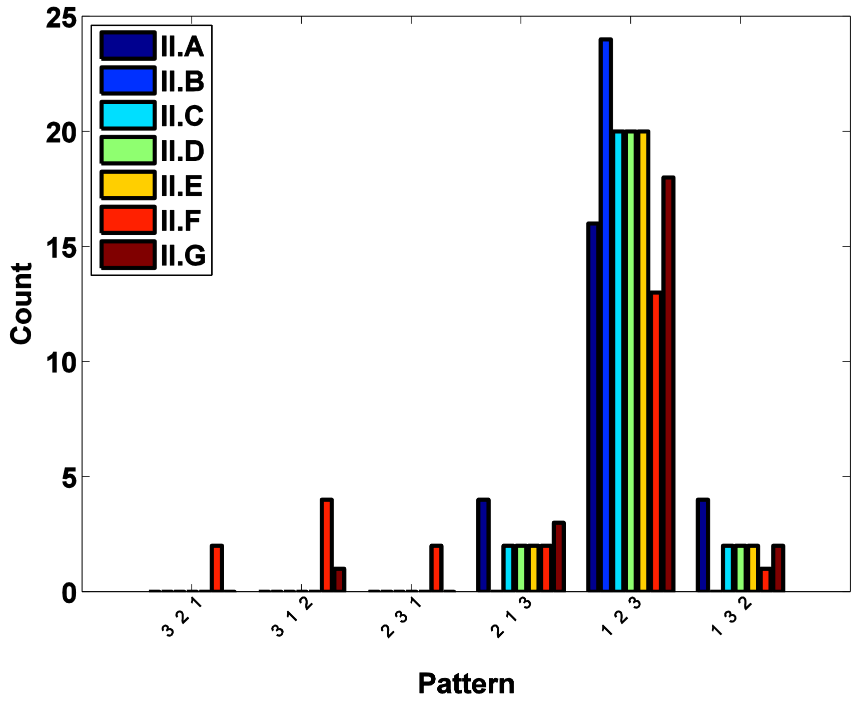

Fig.3b Pattern counts in various time series of Fig.2.

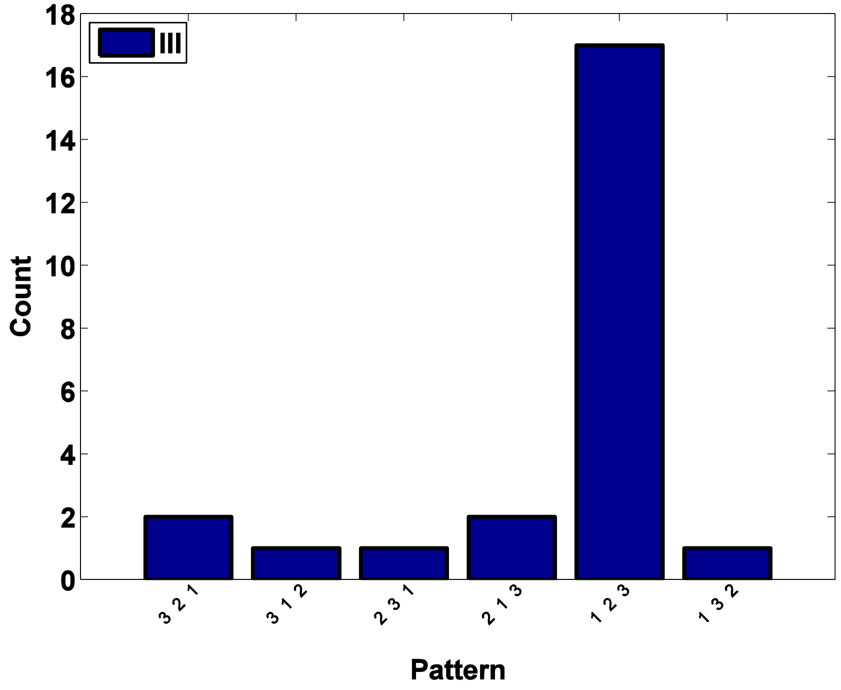

Fig.3c Pattern counts in various time series of Fig.2.

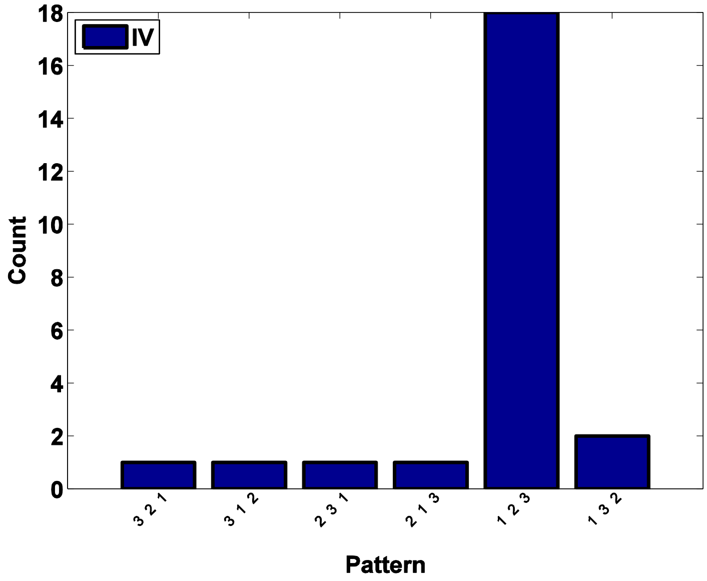

Fig.3d Pattern counts in various time series of Fig.2

Table:1

S.No.	Time series	P_E
1	IA	0.344598
2	IB	1.270228
3	II-A	0.867563
4	II-B	0
5	II-C	0.566086
6	II-D	0.566086
7	II-E	0.566086
8	II-F	1.38437
9	II-G	0.815186
10	III	1.05567
11	IV	0.952513

It is interesting to note that the time series data for export of Chemicals and related products (II-B) shows deterministic trend as the permutation entropy, $P_E = 0$.In other words the system associated with the export of this item is deterministic. One may thus infer that the linked system of policies, agencies and factors associated with Chemicals and allied products in India and abroad from production to markets has evolved into a deterministic complex system in the sense that instruments or intervention for ‘repair and maintenance’ of it are in place though may not apparent but inprinciple decipherable. On a scale of industrial production technology, Chemicals and related products occupy a higher place compared to other manufactured items. One can hypothesise that system associated with export of high technology product may show low complexity index.

Other items in the manufacturing sector namely engineering goods (II-C), textile and textile products (II-D), Gems and jewellery (II-E) show the same value of P_E(note the similar pattern statistics in Fig3b). Though the complexity indices of systems associated with these products is not zero i.e. these systems are not as deterministic as that of Chemicals and related products, it can be inferred that control or regulatory mechanisms inherent in these systems have similar effectiveness.

The highest complexity index is found in thesystem associated with export of handicrafts (excluding hand made carpets (II-F). Clearly this system has not as yet evolved effective control and regulatory mechanisms. Without this its stability remains a question mark.

On the basis of the above results one can conjecture that amongst the manufactured items as Chemicals and related products occupy higher place on technology sophistication scale, the linked system from production to markets in India and abroad has evolved in a stable regime; repair and maintenance (control and regulatory mechanisms) operate to keep the system going in heterogeneous and changing socio-economic contexts.

On the end, handicrafts being at the lowest level on technological scale is long way from evolving repair and maintenance mechanisms perhaps because of strong 'local' or contextual nature of the product that, unlike standardised and mechanised (industrial) mass produced items that get associated with large complex but stable production and market system across socio-economic contexts, gets confined to it own small system.

4.Analysis of Research Publication data in astronomy and Astrophysics

In this section we analyze the research publication data in the area of astronomy and astrophysics during the period 1950-2005. Further the data regarding number of papers, single , double , triple and more than three authors were compiled from the SAO/NASA data system, hosted by High energy astrophysics division of Harvard Smithsonian Centre for Astrophysics.

In Fig.4 (a)-(f), we use the pie- charts to represent the average of the publication data for every 9yrs. We observe that the research publication in the period 1950-1959, due to single author was 89% whereas as it has reduced to 13% during the period 2000-05. Further it is interesting to find that percentage of double authorship article increased from $< 1\%$ (during 1950-59) to a peak of 32% (during 1970-1989) and subsequently decreased to 25% during 2000-05.

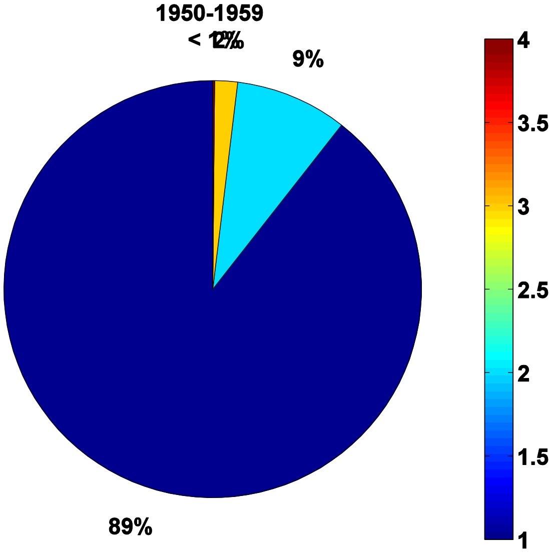

Fig.4 (a) Pie chart for the research publication by single (dark blue), double (light blue), triple (yellow) and more than 3 authors (red) in Monthly notices of Royal Astronomical Society.

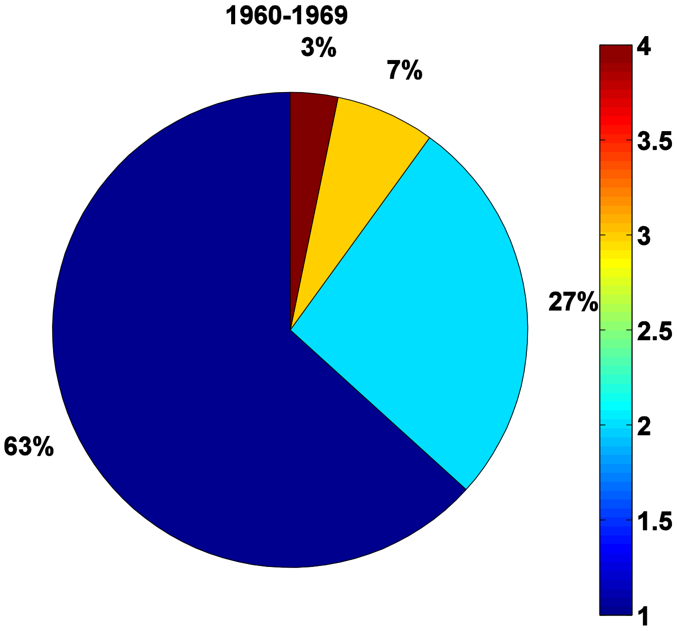

Fig.4 (b) Pie chart for the research publication by single (dark blue), double (light blue), triple (yellow) and more than 3 authors (red) in Monthly notices of Royal Astronomical Society.

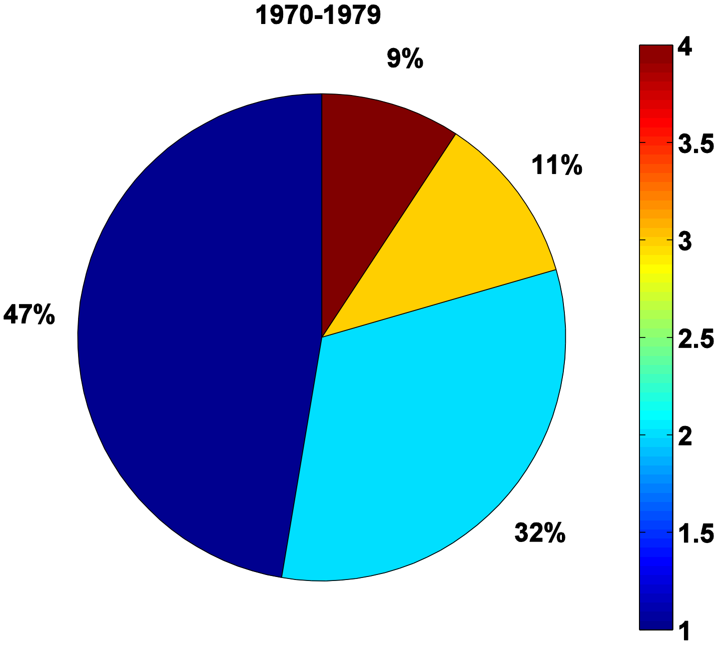

Fig.4(c) Pie chart for the research publication by single (dark blue), double (light blue), triple (yellow) and more than 3 authors (red) in Monthly notices of Royal Astronomical Society.

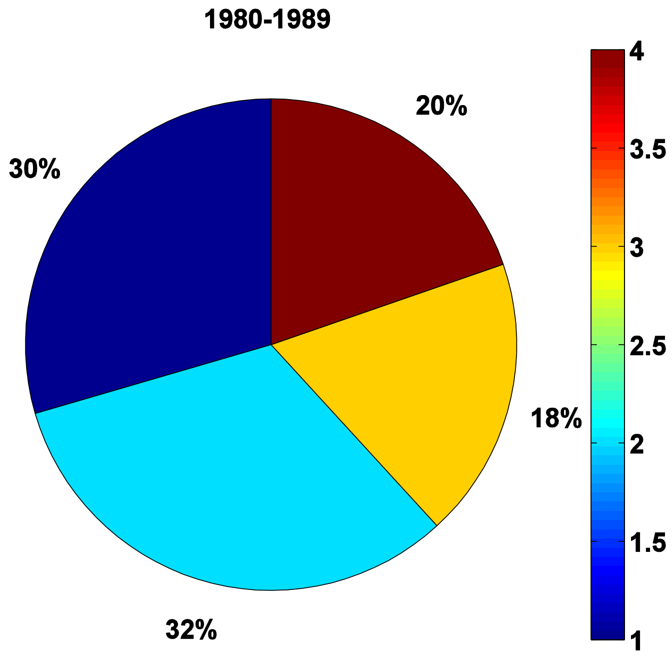

Fig.4 (d) Pie chart for the research publication by single (dark blue), double (light blue), triple (yellow) and more than 3 authors (red) in Monthly notices of Royal Astronomical Society.

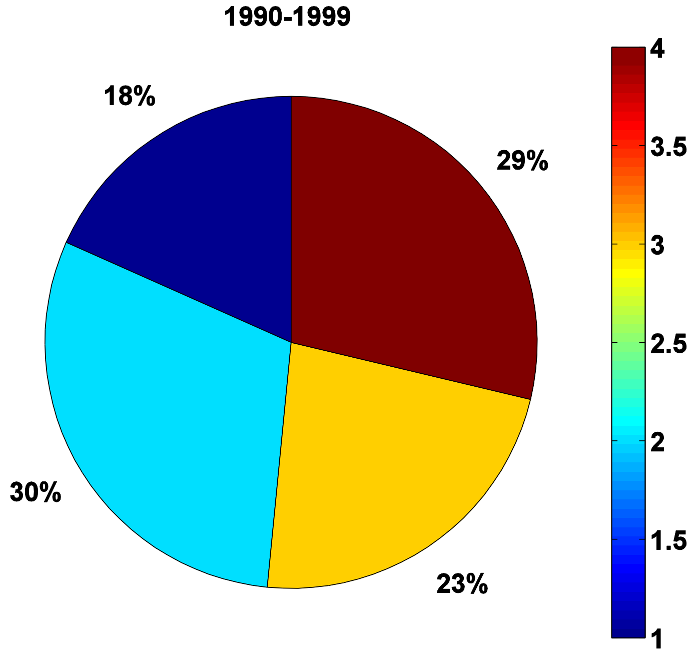

Fig.4 (e) Pie chart for the research publication by single (dark blue), double (light blue), triple (yellow) and more than 3 authors (red) in Monthly notices of Royal Astronomical Society.

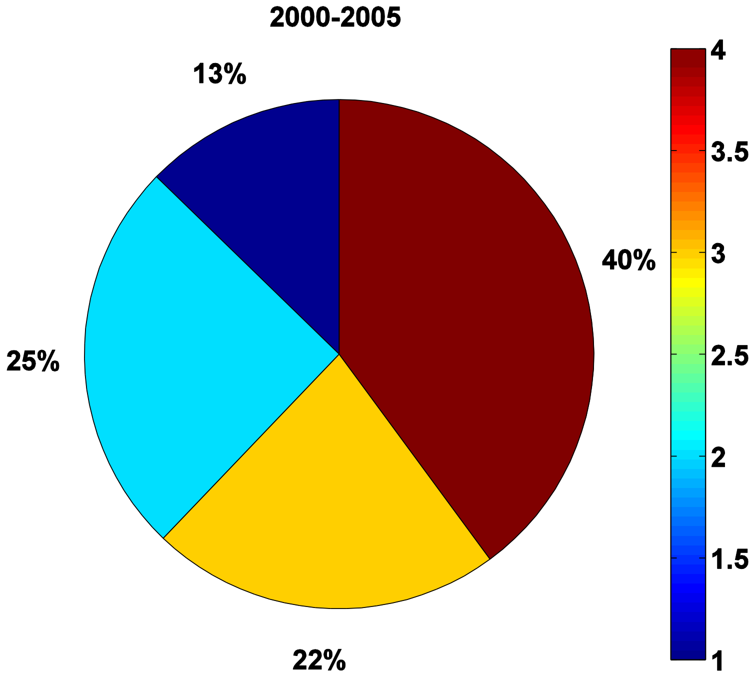

Fig.4 (f) Pie chart for the research publication by single (dark blue), double (light blue), triple (yellow) and more than 3 authors (red) in Monthly notices of Royal Astronomical Society.

The pie- chart of research publications in Astrophysical journal is shown in Fig.5 (a-(f). In Astrophysical journal, on taking the average publication data every 9 yrs, we observe that publication of research paper due to single author has systematically declined from 64% during 1950-59 to 15% during 1990-99. During the period 2000-05, it has reduced to about 5% level. Further the research publication due to double author has shows oscillatory trend i.e. 27% (50-59) to 31% (60-69) to 24%(70-79) to 34%(80-89) to 29%(90-99). There is however systematic increasing trend of research publications by more than three authors.

The time series for research publications in both the journal are shown in Fig.6 (a) and Fig.6 (b). The trend of number of research publication due to single, double, triple and more than 3 authors differs considerably. In a similar way the trend of the total number of publication year-wise in these two journals differs significantly(Fig.7 (a)-(b)).

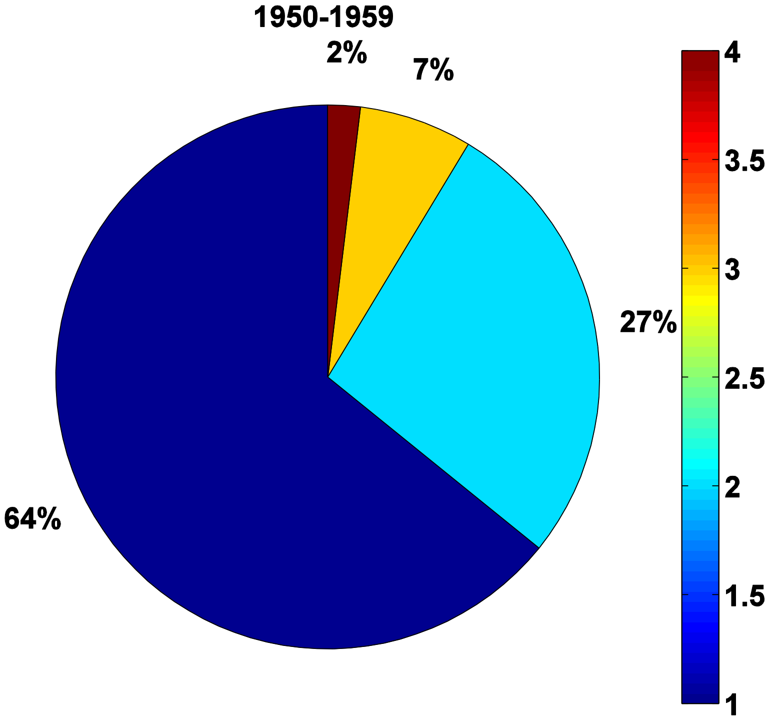

Fig.5 (a) Pie chart for the research publication by single (dark blue), double (light blue), triple (yellow) and more than 3 authors (red) in Astrophysical Journal.

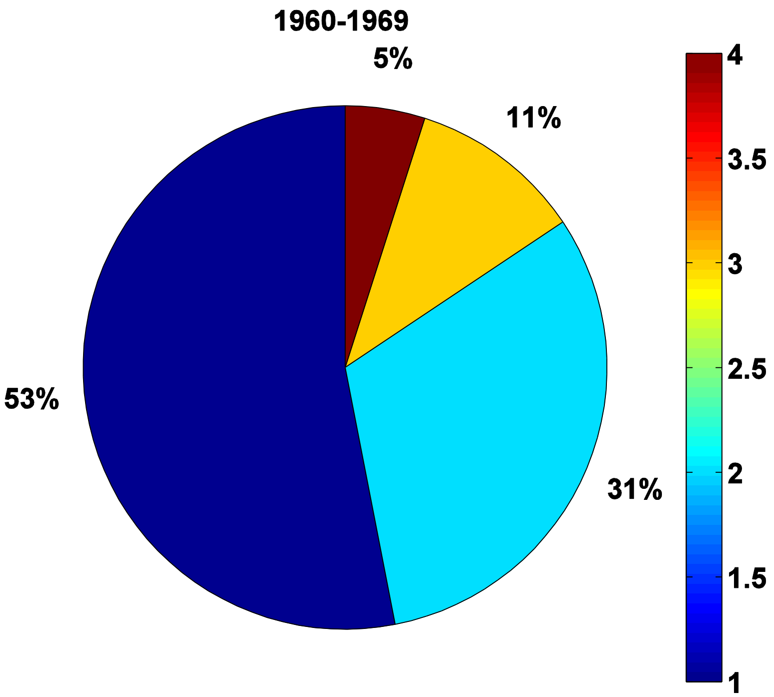

Fig.5 (b) Pie chart for the research publication by single (dark blue), double (light blue), triple (yellow) and more than 3 authors (red) in Astrophysical Journal.

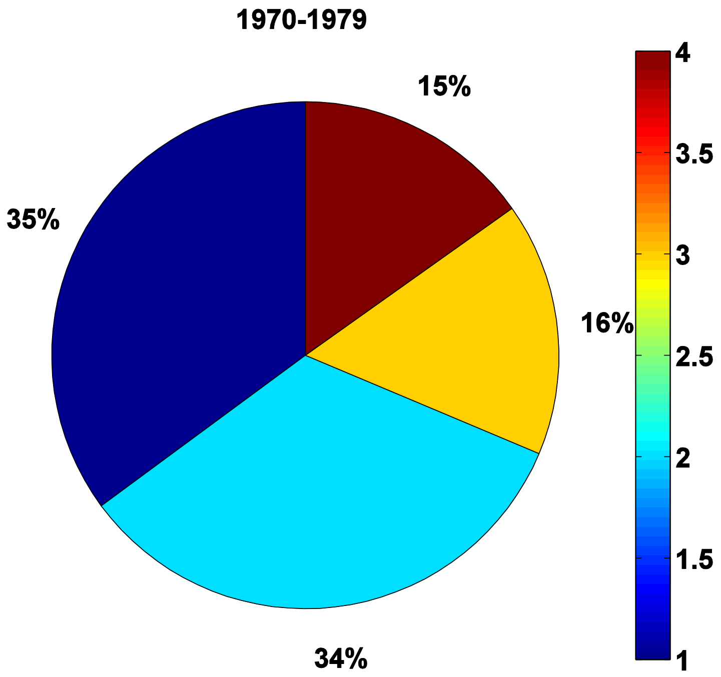

Fig.5(c) Pie chart for the research publicationin Astrophysical Journal: single (dark blue), double (light blue), triple (yellow) and more than 3 authors (red)

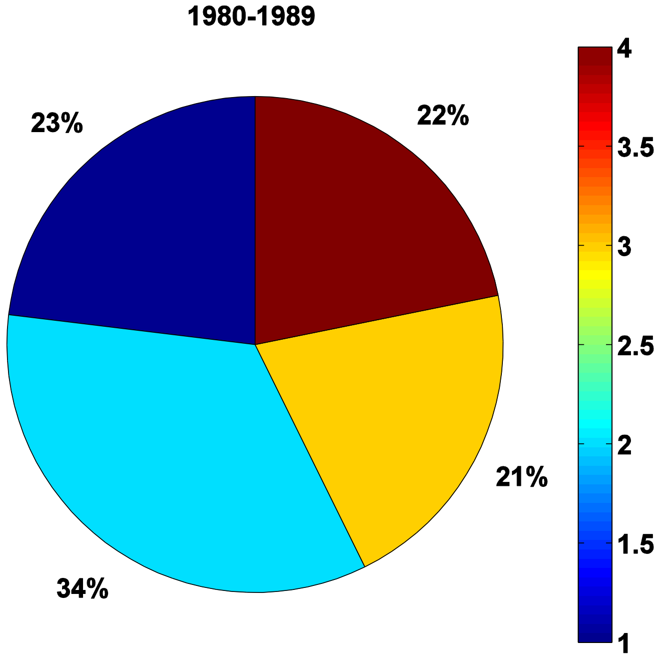

Fig.5 (d) Pie chart for the research publication by single (dark blue), double (light blue), triple (yellow) and more than 3 authors (red) in Astrophysical Journal

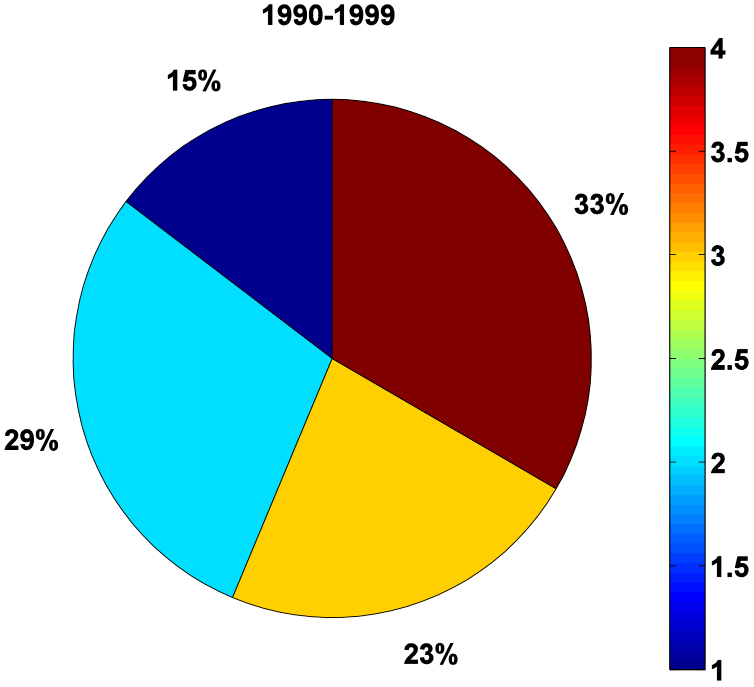

Fig.5 (e) Pie chart for the research publication by single (dark blue), double (light blue), triple (yellow) and more than 3 authors (red) in Astrophysical Journal.

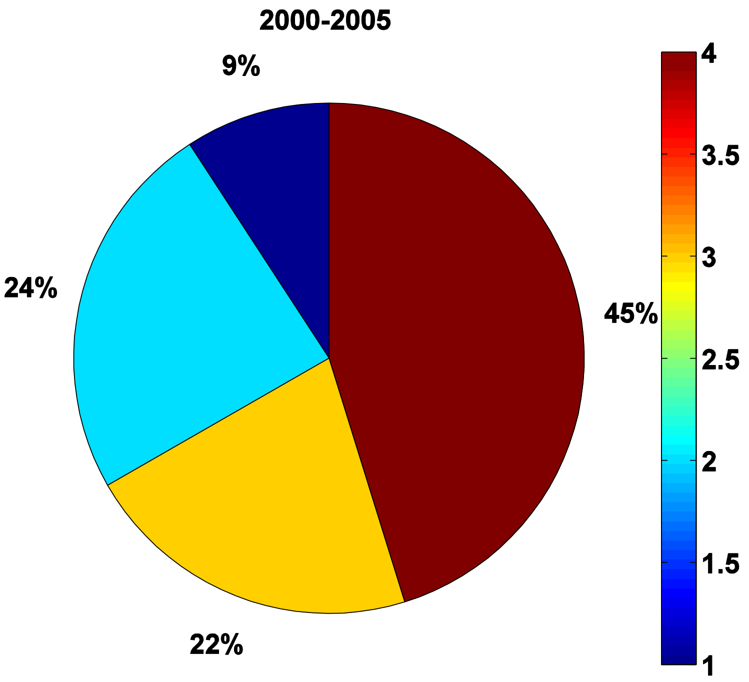

Fig.5 (f) Pie chart for the research publication by single (dark blue), double (light blue), triple (yellow) and more than 3 authors (red) in Astrophysical Journal.

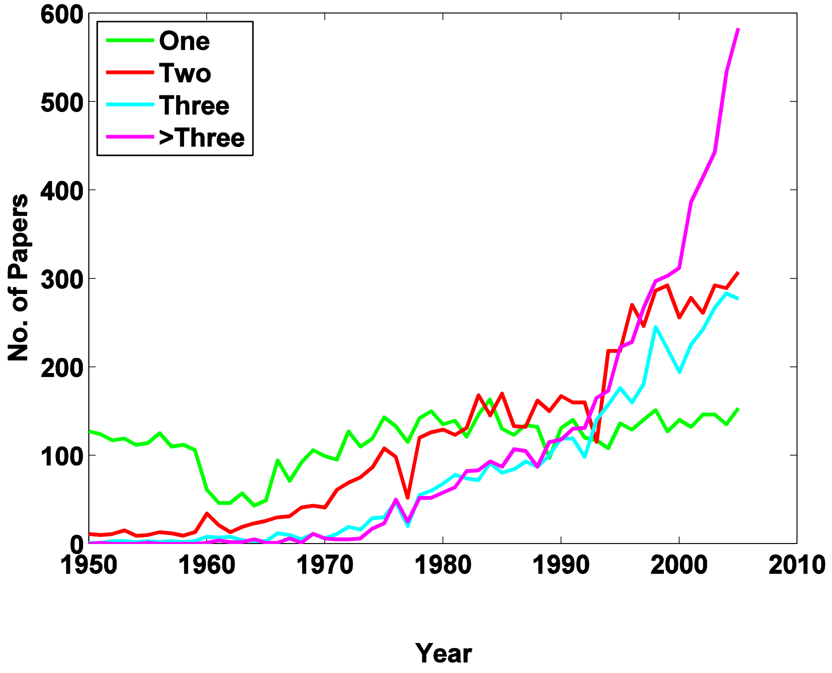

Fig.6(a), Time series data of research publication due to single, double , triple and more than 3 authors during 1950-2005 in Monthly Notices of Royal Astronomical Society.

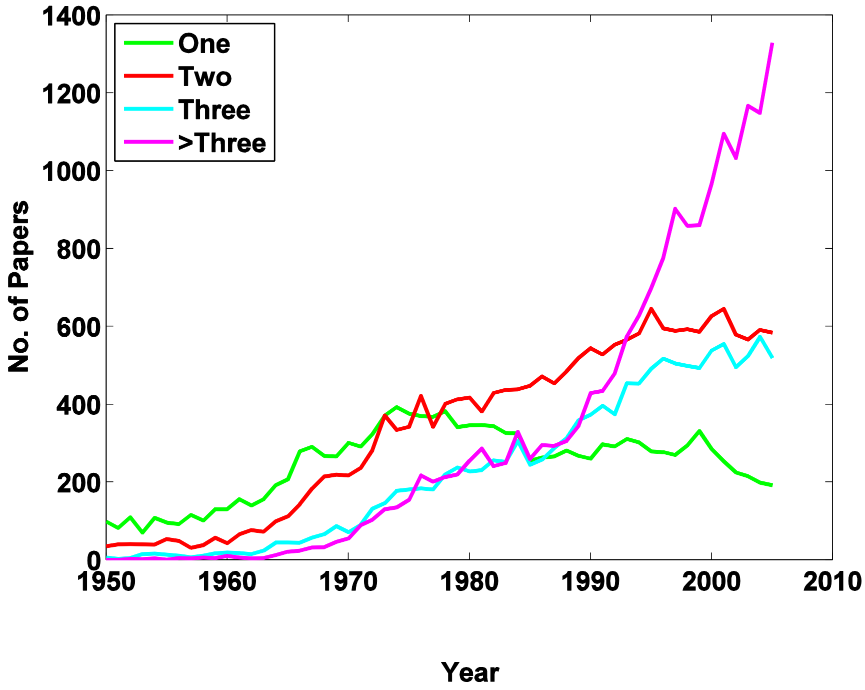

Fig.6(b), Time series data of research publication due to single, double , triple and more than 3 authors during 1950-2005 in The Astrophysical Journal.

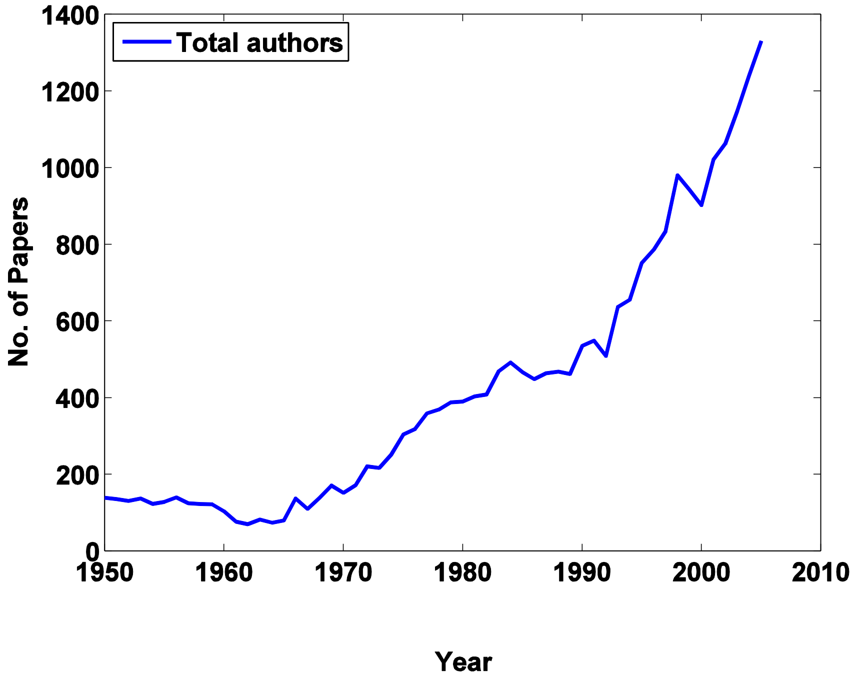

Fig.7 (a). Time series data of totalresearch publication during 1950-2005 in Monthly Notices of Royal Astronomical Society.

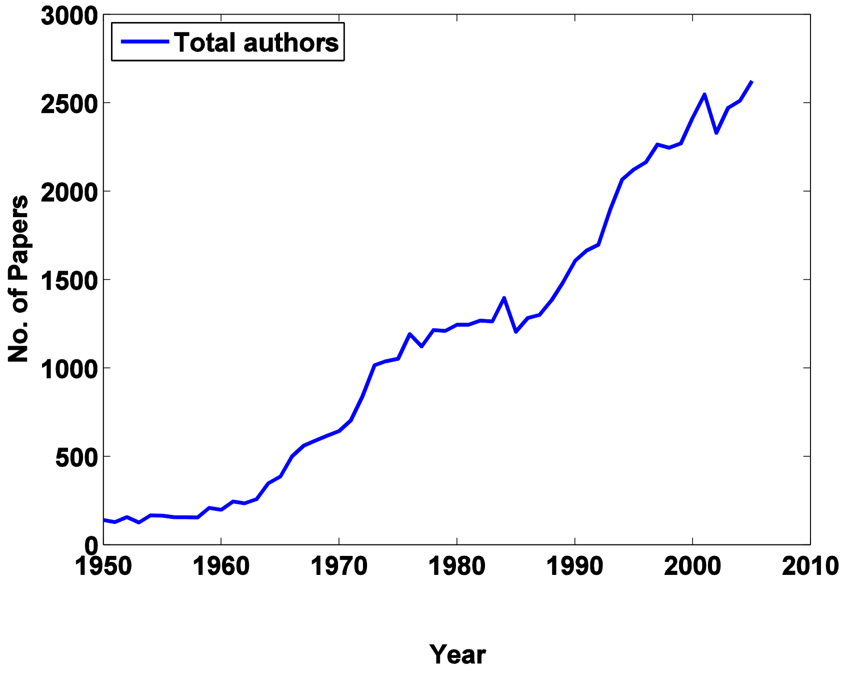

Fig.7 (b), Time series data of total research publication during 1950-2005 in The Astrophysical Journal.

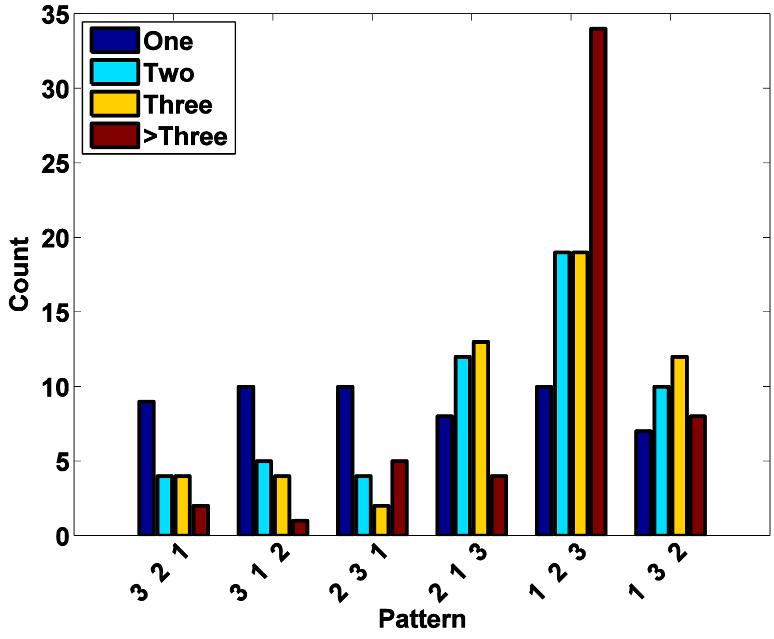

Fig.8(a) Pattern statistics in time series data of research publication due to single, double , triple and more than 3 authors during 1950-2005 in Monthly Notices of Royal Astronomical Society.

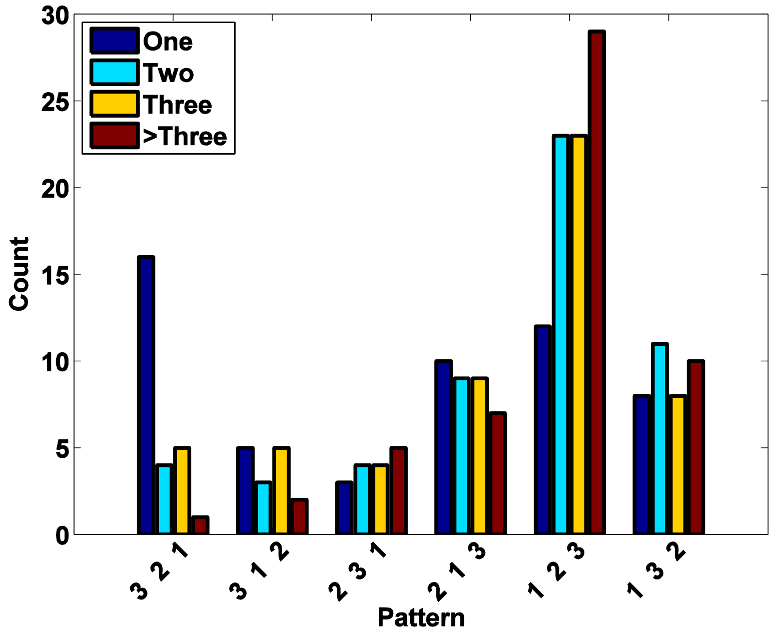

Fig.8(b) Pattern statistics in time series data of research publication due to single, double , triple and more than 3 authors during 1950-2005 in The Astrophysical Journal.

We have also carried out the computation of various pattern statistics of publication in these two journals for the period 1950-2005. These bars chart also differs considerably and are shown in Fig.8 (a)-(b). These show possibility of different dynamics of publication in this journal.

The permutation entropy [8] computed for research publication data in the above two journals are shown in Table: 2. It is observed that the time series of single and double author publication in MNRAS (Monthly Notices Royal astronomical Society) is more complex than single and double author publication time series in APJ(The Astrophysical journal). Similar result follows for the time series of total publications. Further in case of more than three authors, the complexity is more in APJ than MNRAS. These results clearly show the existence of differentmechanism in operation for publication in the two journals.

Table: 2 Permutation entropy,P_E

S.No.	Journal	Single author	Double author	Triple author	More than 3 authors	Total No. of papers
1	APJ	1.670749	1.532421	1.578494	1.327275	1.317256
2	MNRAS	1.783253	1.619973	1.552239	1.183235	1.523555

It is observed that relatively, the complexity ublication

5.Conclusion:

It is shown that fluctuations in time series can be used to measure the nature of complexity in processes embedded in economic or knowledge generating activity in terms of complexity index. It is emphasized that the inferences drawn from complexity measures are based purely on observed fluctuations in data series and do not invoke any prescribed theoretical models or assumptions regarding either factors involved in the economic or knowledge activity under consideration nor about correlations between them. The approach is emerging as a technique of extracting information from time series data on processes involved in the economic activity that otherwise remained ignored.

The results of application of technique of Permutation Entropy or Symbolic Time Series Analysis (STSA) to time series data of exports presented in the paper prompts one to hypothesise that complexity index may be another method of grading exported products on technology sophistication scale. Higher the index lower would be the position on technological sophistication scale. As part of a larger project, the authors propose to further develop and test the hypothesis by using time series data of exports from other countries of products classified in literature as belonging to ‘high’ or ‘low’ technology.

It is also expected that the emerging new taxonomy based estimation of complexity index from time series data in economics and research publications will inform a) design of specific intervention instruments to direct and analyse the performance of activities involved and b) provide additional perspective to theory based modelling work

Acknowledgements

Our foremost acknowledgement is to Knowledge Forum, the interdisciplinary nature of which is the main trigger for this work and to Prof Siddharthan for suggestions relating to interpretation of findings. Thanks are due to Dr.Naveen Bhatraju of IGIB, New Delhi for providing help in computation of permutation entropy and Mr.Sharad Mishra in making diagrams, figures and formatting the paper.

References

[1] Arun Maira, Redesigning the Aeroplane while Flying - Reforming Institutions, Rupa Publications Pvt. Ltd, 2014, pages 66-67.

[2] Juan Gabriel Brida,Structural Change and Economic Dynamics, Vol. 14, No. 2, pp. 159-183, October 1, 2000.Accessed on July 02, 2014 from decon.edu.uy/publica/Doc1000.pdf.

[3]Takuya Yamamoto, Kodia Sato, Taisei Kaizoji and Jan-Michael Rost, Symbolic analysis of indicator time series by quantitative sequence alignment, Journal: Computational Statistics & Data Analysis, Volume 53, Issue 2, December, 2008, pages 486-495. Elsevier Science Publishers, the Netherlands.

[4] Daw, C.S., Finny, C.E.A., and Tracy, E.R., 2002, A review of symbolic analysis of experimental data, http:/ www-chaos.engr.utk.edu/pap/crag –rsi2002.pdf.

[5] Bandt, C., Pompe, B : 2002, Phys.Rev.Lett., vol 88,174102.

[6] Da-Guan Ke and Qin-Ye Tong, 2008, Easily Adaptable Complexity measures for Finite Time Series, Phys.Rev.E. 77,066215.

[7] Piccardi, C., 2006, On Parameter estimation of chaotic systems via symbolic time-series analysis, CHAOS, vol. 16, 043115.

[8] Jain, A., and Das, M.K., 2013, Complexity Measure in Publication data, National Workshop on “Measuring science: The Scientometric Approach”, NISTADS.