Production and productivity of Chinese scientists as a function of their age: the period 1995-1999

Bihui Jin[1], Ling Li[1], Ronald Rousseau[2]

[1] Documentation and Information Centre of the Chinese Academy of Sciences
Chinese Science Citation Database
33 Beisihuan Xilu, Zhongguancun, Beijing, 100080, P.R. China
e-mail: jinbh@mail.las.ac.cn

[2] KHBO, Dep. Industrial Sciences and Technology
Zeedijk 101, B-8400 Oostende, Belgium
&
LUC, Universitaire Campus,
B-3590 Diepenbeek, Belgium
e-mail: ronald.rousseau@khbo.be

Abstract

Based on the 1995-1999 data from the Chinese Science Citation database (CSCD) this article investigates the year-by-year age distribution of scientific and technological personnel publishing in China. It is shown that the 'Talent Fault' originating during the Cultural Revolution still exists, and that a new gap resulting from recent brain drain might be developing. The purpose of this work is to provide necessary information about the current situation and especially the existing problems of the S&T workforce. We hope that our discussion will help science policy makers to make more adequate use of existing talent by restructuring and optimizing the scientific environment. Such an environment should foster a contingency of talent fitting in the highly competitive environment of the new century.

Keywords: age structure, scientific and technological talent, research policy, talent gap, productivity

1. Introduction

Intellectual and technological talents and skills are the driving force for scientific and industrial development. A nation's research capacity and its ability to transfer research achievements to other sectors of society are closely related to its knowledge structure. An important aspect of this knowledge structure is the age distribution of a country's S&T workers. Consequently, also in China, this age structure has been given close attention by administrative departments. In 1985, the Chinese Academy of Sciences completed statistics on the age structure of those scientists who were responsible for a total of 1270 research projects. Hu Xiaoyuan et al. (1993) used a questionnaire approach in order to obtain statistics on the age of 2426 authors of research papers. Despite the drastic difference in methods and objects, these two investigations lead to similar results, indicating a clear gap in the use of the

Bihui Jin, Ling Li and Ronald Rousseau, *Production and productivity of Chinese scientists as a function of their age: the period 1995-1999*, In: Guohua Jiang, Ronald Rousseau, Yishan Wu (Eds), Proceedings of the 9th International Conference on Scientometrics and Informetrics – ISSI 2003, Dalian University of Technology, Dalian, China, 2003, 112-120.

country's inherent talent. The 1985 statistics show a gap between the age of 36 and 40 years and Hu's statistics, collected about eight years later demonstrate that the gap had widened to the age span between 32 and 47. According to Hu et al. (1993), the publication productivity peak was situated between 49 and 58 years, 8 years later than the result of 1985.

The reason for this gap is well-known. It is the result of the Cultural Revolution, when people with the inherent talents to do S&T work were not given the opportunity to study and acquire the necessary skills. Recently, Liang Liming et al. (2001) and Wu Yishan et al. (2003), confirmed the continuing existence of this gap in the use of China's latent talents.

Based on the 1995-1999 data from the Chinese Science Citation database (CSCD) and by using a bibliometric method, this work investigates the year-by-year age distribution of scientific and technological personnel publishing in China. The purpose of this work is to provide necessary information about the current situation and especially the existing problems of the S&T workforce. We hope that our discussion will help science policy makers using existing talent more adequately by restructuring and optimizing the scientific environment. Such an environment should foster a contingency of talent fitting in the highly competitive environment of the new century.

2. The general age distribution of scientists and their production

According to Zhao's theory of the *Optimal Age for Scientific Creativity* (Zhao, 1984) man's creativity is at its highest between the ages of 25 and 45, reaching a peak around 37 years. Consequently, if this theory is correct, universities should aim at having most active scientists in that age group.

The relation between productivity and age has been studied by many researchers. Kyvik (1990) studied the relationship between age and scientific productivity in Norway. He found that publishing activity reaches a peak in the 45-49 year old age group and declined by 30 percent among researches over 60 years old. He, moreover, found that large differences existed between different fields of learning. In the social sciences, for instance, productivity remains more or less at the same level for all age groups, while in the natural sciences productivity continually decreases with increasing age. This finding corresponds with Diamond's (1986) observation that for mathematicians quantity and quality of output declined monotonically with age. Total citations peaks, however, came much later.

Zusne (1976), plotting the age-group at which the most significant publication of a group of 213 psychologists took place, found a curve closely following Zhao's (1984) and Lehman's (1953) suggestions, with a top in the age group between 35 and 40 years of age. Note, however, that there is a big difference, between plotting one top performance (usually an act of high creativity) or the yearly number of publications. The latter number depends much more on perseverance and acquired power.

It has been observed (Kyvik, 1990) that collaborative research can sustain productivity when researchers become older. Differences between the peak performances in different fields, and this over several centuries, were also studied by Liang et al. (1999). These authors even found that, generally, the peak ages for the sciences and for the arts were similar.

Melker (1999), while studying the situation in Russia, found two age distributions for scientists: one for the Academy and industrial institutes, and one for universities. The first one has a bell shape, peaking around the age of 60, the latter one has two peaks: a big one around the age of 25, and a smaller one around the age of 60. He calls the first one a dromedary curve (a situation leading to stagnation and then the death of science), so that – adhering to his analogy – the second situation could be called a camel curve.

3. Source of the statistics and basic data

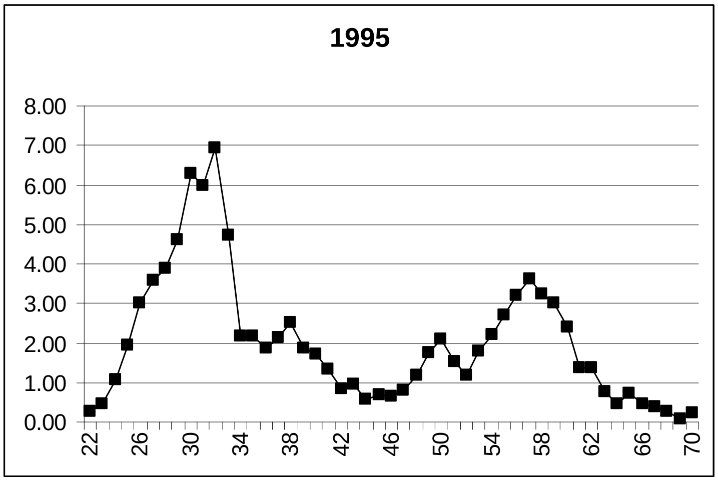
The used data are obtained from the CSCD, a database constructed by the Documentation and Information Centre of the Chinese Academy of Sciences, and supported by the National Natural Science Foundation of China and the Chinese Academy of Sciences. It covers more than 600 important journals published in China, including those covered by SCI-Expanded. For more information on this database and its uses the reader is referred to the following articles, where more details are provided (Jin & Wang, 1999; Rousseau et al., 2001; Jin & Rousseau, 2001; Jin et al., 2002). Besides standard bibliographical information, the CSCD also processes, since 1995, authors' information such as professional title, gender, and age. Consequently, the authors were able to use altogether 131,615 records of authors' ages for the period 1995-1999.

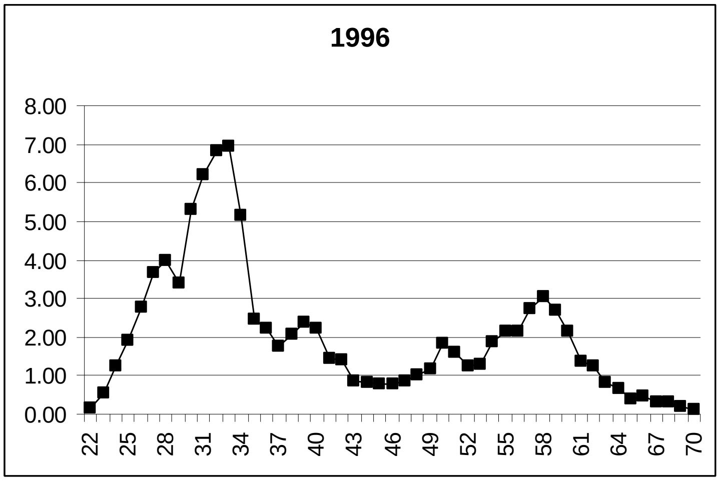
Not all journals, however, provide age data. In 1999, for example, only 487 of the 633 source journals included authors' ages. Moreover, in most cases, only the first author's age is given. Hence, this study is for the larger part based on first author's ages. Only occasionally the ages of other authors are taken into account. According to a recent study by Liang Liming et al. (2001) this means that our results are based on an underestimate of older scientists, as they often occur last in the authors' byline (at least in the computer sciences, but probably also in many other fields). In China ranking of authors generally conveys the contribution of the co-authors. For this reason we may say that the ages of the first authors are representative for the most active scientists.

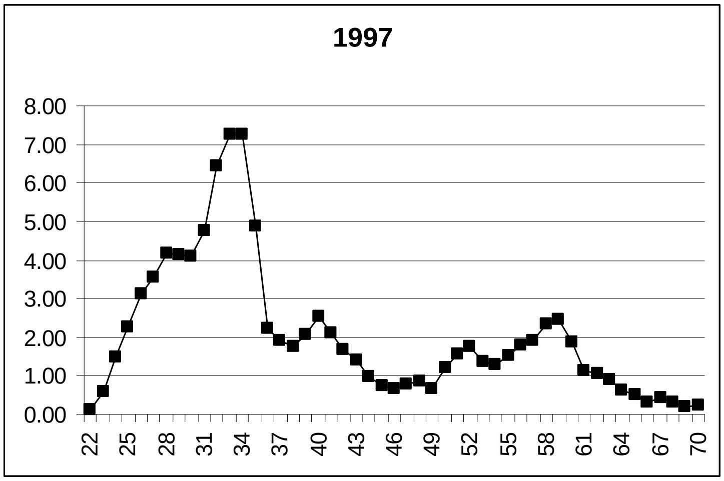
Based on 131,615 records we calculated, for each year of the 1995-1999 period, the proportion of the papers authored by people of each age class, leading to a list of essential data on the age distribution of first authors, graphically shown in Fig.1.

4. Visible characteristics of yearly authors' age distributions

Fig.1 shows the complete age distributions for the five analysed years, with authors' age on the abscissa and percentages on the ordinate. Because of the wide age span (from 18 to 90 years) we will henceforth leave ages under 22 and above 70 out of consideration. The general shape of these curves coincides with a 'camel' curve. In order to better display the overall development the curves for the years 1995 and 1999 are shown together in Fig.2. In Jin et al. (2002) we showed that we may restrict ourselves to first authors' curves (by comparing them to all authors' curves).


1995
8.00
7.00
6.00
5.00
4.00
3.00
2.00
1.00
0.00
22
26
30
34
38
42
46
50
54
58
62
66
70


1996
8.00
7.00
6.00
5.00
4.00
3.00
2.00
1.00
0.00
22
25
28
31
34
37
40
43
46
49
52
55
58
61
64
67
70


1997
8.00
7.00
6.00
5.00
4.00
3.00
2.00
1.00
0.00
22
25
28
31
34
37
40
43
46
49
52
55
58
61
64
67
70

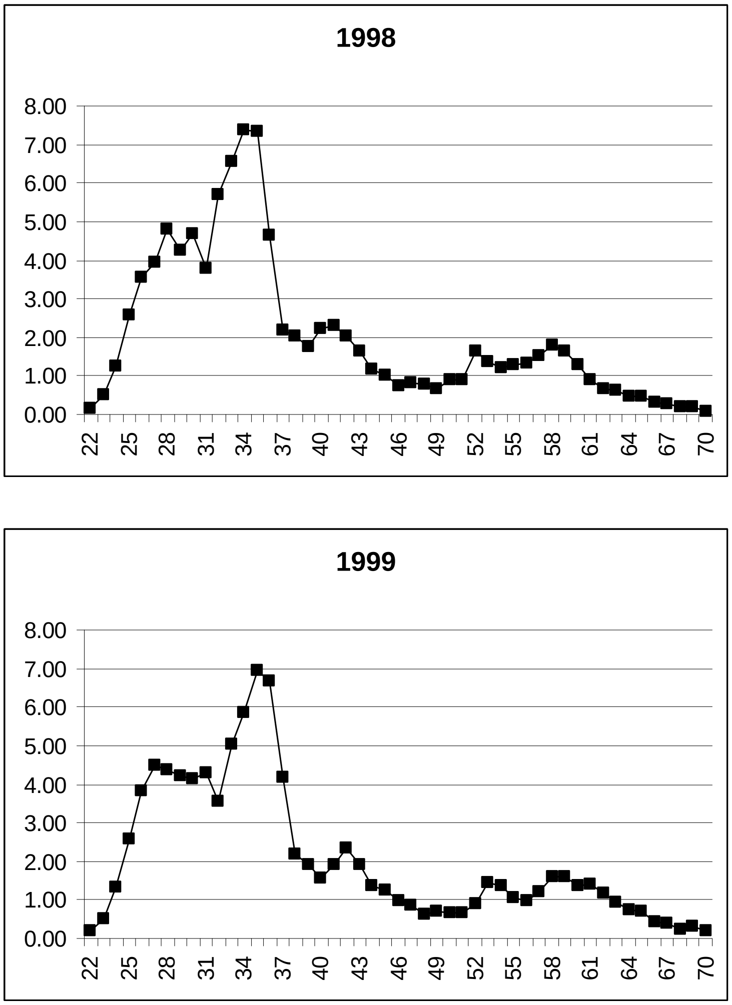


Fig.1 Year by year evolution of the age distribution of active Chinese scientists

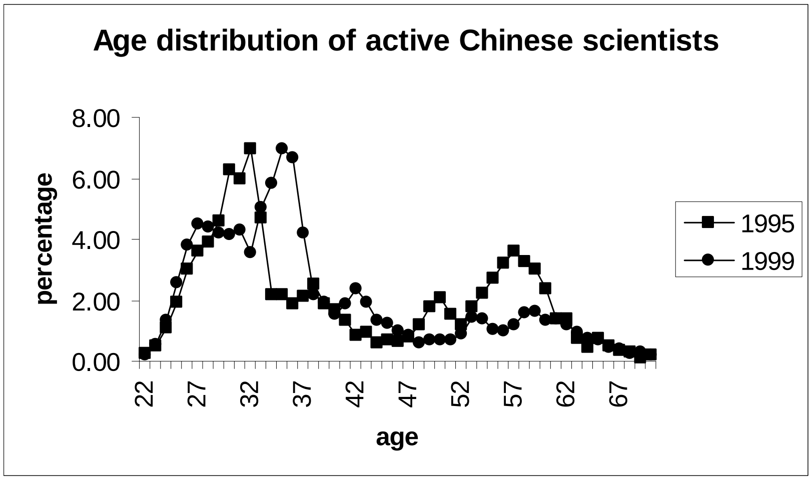


Fig.2 A comparison between the years 1995 and 1999

Comparing the curves for the years 1995 and 1999 we notice that the largest peak corresponds to younger authors (35-36 years old in 1999). These scientists received their training in recent years and are probably publishing the results of their doctoral theses. This peak moves to the right, and hence refers roughly to the same group of persons. There is another peak situated around 57 years in 1995. This peak diminishes over the years. It corresponds to older scientists who received their training before the Cultural Revolution (which began in 1966). There is a clear gap corresponding to those scientists who should have received scientific training during the period 1966-1976, but, as is well known, did not receive any education. Only a minority of these became later active scientists. This period corresponds to a period where China wasted the S&T talent it had. Between the two major peaks and the gap due to the Cultural Revolution, we notice several minor peaks and gaps indicating a – probably natural – alteration of small ups and downs.

One could say that, at least in general terms, the age distribution curve approaches Zhao's ideal curve. Indeed, in 1979 the Chinese scientometrician Zhao Hongzhou proposed that the optimal period of scientific creativity is between 25 and 45 years, with a top around 37. Nowadays, in the era of Big Science, is seems likely that, maybe not creativity, but certainly productivity, reaches a peak later on in life.

Finally, if we look carefully, we may notice a disturbing trend among the young scientists. On the 1995 curve we see that the increase in production slows down somewhat around the age of 29. On the 1999 curve this is not just a small disturbance anymore, but we notice a widening plateau from 29 to 33. It even seems that a new valley is forming. Why is this happening? Nowadays many young scientists and technicians showing great promise prefer to go abroad to study (and often stay there), or start working for hi-tech companies (abroad or in China). In all these cases there is a loss for science, and a possible brain drain for the country.

5. Three different age groups

Fig.3 shows the relative production of papers (over the five years studied) depending on the year of birth.

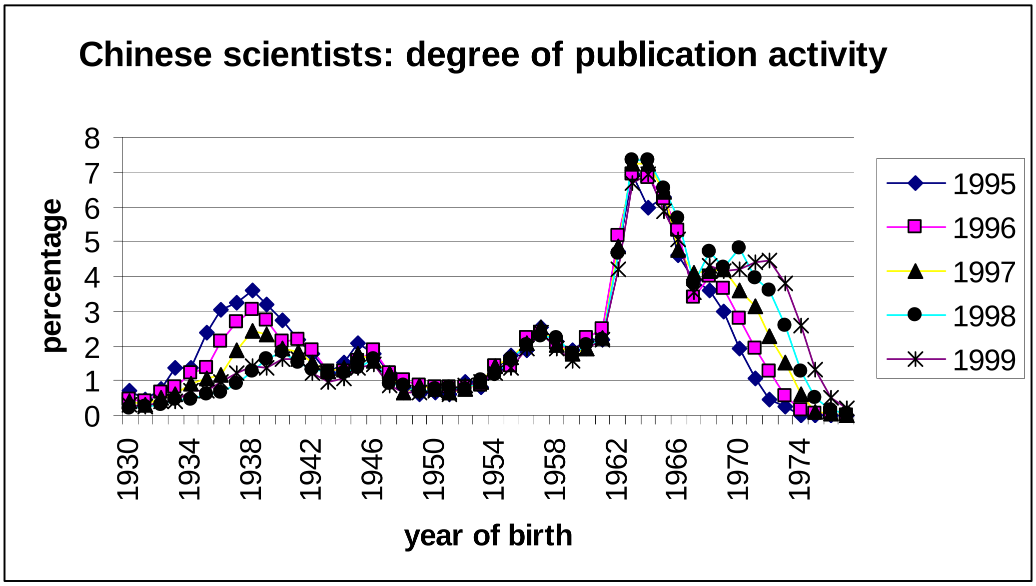


Fig. 3 Relative production of papers depending on the year of birth of the first author

In this picture we may distinguish three groups of authors. First there is the group of scientists born after 1968. These authors have an increasing share in China's scientific production. Yet, careful examination shows that the year 1971 is a turning point. Authors born after 1971 show a fast increase, while those before 1971 a much slower increase (an even a decline in 1999).

Second there is the stable group of authors born between 1943 and 1967. Among these authors we notice the group of high-production and the small group of scientists who keep on working despite the hardship they encountered during the Cultural Revolution.

Finally we see that the authors born between 1933 and 1942 decrease their share in the total scientific production. These scientists are retired or about to retire, so that their production decreases in a natural way. It is, however, their dedication to scientific research that still contributes to the existence and dynamic change of a peak in this age rank.

6. A mathematical analysis of the age curves

In a stable situation publication age curves look the same each year. This is clearly not the case with the Chinese publication age curves, as the major peak moves in time, coinciding with the same cohort of scientists. We will now show how statistical techniques can reveal these features.

Similarity between time series and the time shift that produces the greatest similarity can be ascertained quantitatively using the notion of cross correlation (Shiavi, 1991). We observe that a time series, X(n), $n \in \mathbb{N}$, is a discrete stochastic signal. We use the term 'stochastic' because its actual values depend on chance. We will always assume that the stochastic time signals we will consider are wide sense stationary (WSS), and when considering two signals together we will assume that they are jointly WSS (Papoulis, 1991, p.298). As with any other stationary stochastic data set, one can calculate the mean (the first moment), denoted as E[X] or m_X, other moments if necessary, and the standard deviation, denoted as σ_X. If signals X(n) and Y(n) are given, where n denotes the number of time units, then the cross correlation function (CCF) of the signals X and Y (in this order), and denoted as R_{XY}, is given as:

$$R_{XY}(k) = E[X(n)Y(n+k)], \qquad k \in \mathbb{N}$$

Similarly, the cross covariance function (CCVF) of X and Y, denoted as C_{XY}, is defined as:

$$\begin{aligned} C_{XY}(k) &= E\left[\left(X(n)-m_X\right)\left(Y(n+k)-m_Y\right)\right] \\ &= R_{XY}(k) - m_X m_Y \end{aligned}$$

Finally, the normalized cross covariance function (NCCF) of X and Y, denoted as ρ_{XY}, is defined as:

$$\rho_{XY}(k) = \frac{C_{XY}(k)}{\sigma_X \, \sigma_Y}$$

Note that C_{XX} is nothing but the variance of X (σ_X^2) and that, consequently, $\rho_{XX}(0) = 1$. If the averages of the two signals are zero, then their cross correlation and cross covariance functions coincide. We calculated the first components of the normalized cross covariance functions ρ_{ij}, where i and j denote the years 95 to 99, for the components k = 0,…,5. (See Tables 2a and 2b, and Fig. 4).

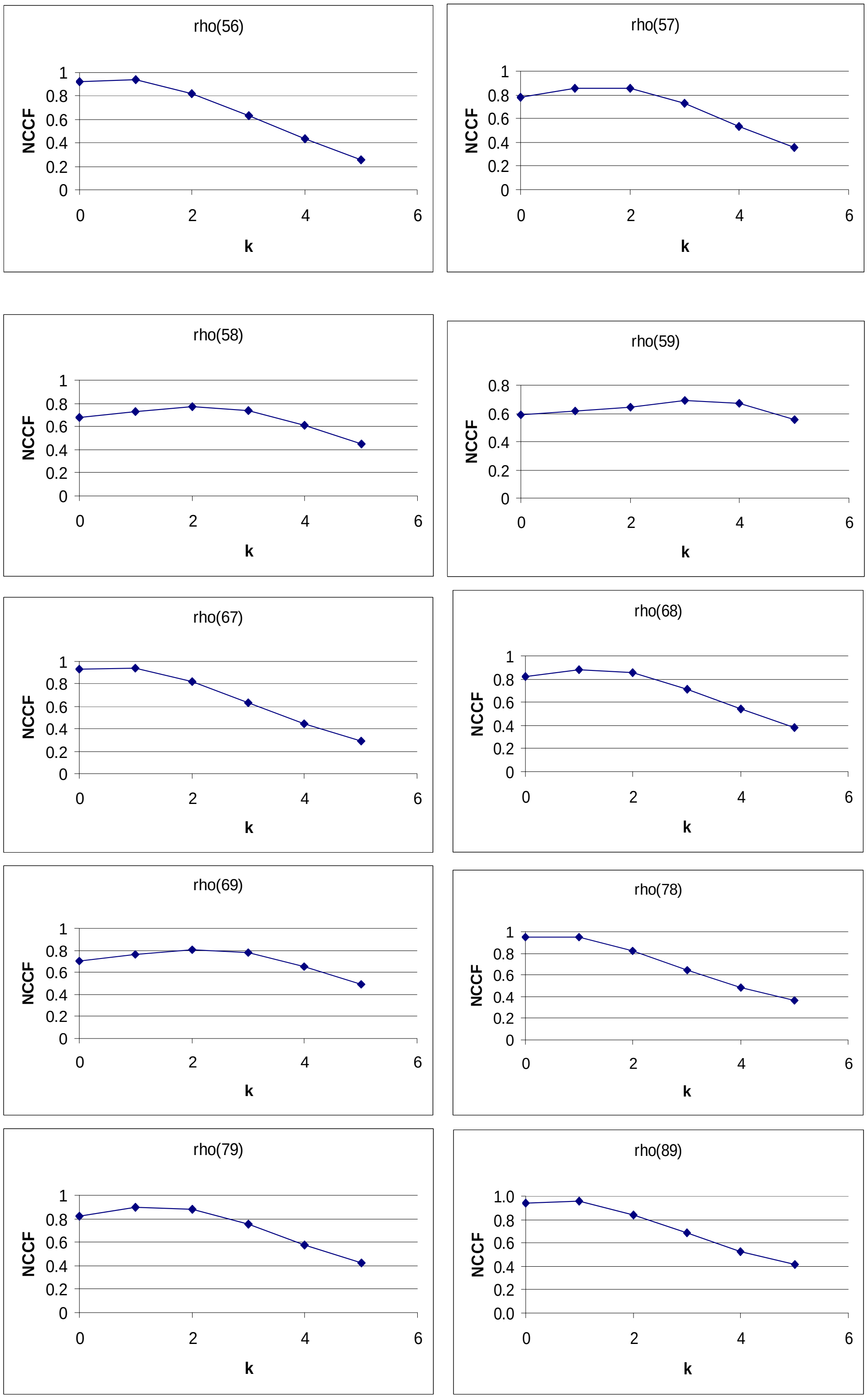


Fig. 4 Normalized cross covariances between every pair of years over the period 1995-1999

Table 2a Cross correlations, continued in Table 2b

k	$\rho_{95,95}$	$\rho_{95,96}$	$\rho_{95,97}$	$\rho_{95,98}$	$\rho_{95,99}$	$\rho_{96,96}$	$\rho_{96,97}$
0	**1.000**	0.919	0.780	0.676	0.588	**1.000**	0.933
1	0.879	**0.942**	**0.859**	0.727	0.617	0.925	**0.944**
2	0.692	0.821	**0.859**	**0.773**	0.646	0.742	0.823
3	0.486	0.633	0.728	0.740	**0.690**	0.548	0.630
4	0.283	0.440	0.537	0.608	0.668	0.370	0.443
5	0.146	0.254	0.354	0.447	0.558	0.237	0.291

Table 2b Cross correlations

k	$\rho_{96,98}$	$\rho_{96,99}$	$\rho_{97,97}$	$\rho_{97,98}$	$\rho_{97,99}$	$\rho_{98,98}$	$\rho_{98,99}$	$\rho_{99,99}$
0	0.824	0.706	**1**	0.945	0.819	**1**	0.937	**1**
1	**0.878**	0.763	0.899	**0.947**	**0.896**	0.904	**0.955**	0.906
2	0.858	**0.801**	0.715	0.826	0.881	0.755	0.843	0.743
3	0.716	0.778	0.535	0.642	0.753	0.591	0.683	0.579
4	0.543	0.649	0.401	0.480	0.577	0.474	0.522	0.463
5	0.382	0.492	0.320	0.364	0.424	0.411	0.417	0.407

Clearly, age curves in subsequent years, shifted over one unit (one year), show a very large similarity. The corresponding NCCF-value is always larger than 0.94. Further, shifting age curves in such a way that scientists born in the same year are compared, always yields large NCCF-values. Even comparing the 1995 curve with the 1999 curve, shifted over 4 units, yields an NCCF-value of 0.7.

7. Discussion

Before discussing the results and distinguishing features of our data we like to make point a technical point. It is sometimes stated, or implied (Bayer & Dutton, 1977) that authors should not be compared based on their age, but on the time since they have entered the field, usually taken as the date of their Ph.D. This is then referred to as *career age*. However, we do not have these data for Chinese authors. We think that our results are interesting enough as they are, without the correction for career age. Using career age would certainly make the gap due to the Cultural Revolution more visible. This, however, would not tell us anything about how a country uses the available scientific brains for the progress of science, according to the population's age.

The S&T age structure shows a dynamic mechanism in movement and evolution. It is governed and shaped by internal and external forces. The internal force is the physiological process of talents, while social factors, political, economic, and cultural influences act as an external force. The external force is the dominant factor when comparing different countries and different years.

When the S&T age structure becomes abnormal due to external events, it takes decades to return to normality, requiring several generations' efforts. The young play a key role in reshaping such an abnormal situation. Considering the S&T system as a market, we conclude that market economy requires new recruiting and motivating mechanisms in order to maintain an adequate contingency of scientific and technological workers. This is vital for the scientific well-being of the country.

Acknowledgement. The authors thank their colleagues Liu Xiaomin, Yang Ninghui, Yuan Yulan, Lu Yun, Zhu Xianyou, Wang Zhinan and Liang Liming for their help during the preparation of this article.

References

Bayer, A.E. & Dutton, J.E. (1977). Career age and research-professional activities of academic scientists. *Journal of Higher Education*, 48(3), 259-282.

Diamond Jr., A.M. (1986). The life-cycle research productivity of mathematicians and scientists. *Journal of Gerontology*, 41, 520-525.

Hu, XY., Cao, XY. & Su JY. (1993). Survey of the fault problem of outstanding talent in China. *Bulletin of Science and Technology in China*, 3, 30-32 (in Chinese).

Jin, BH, Li, L., Liu, XM., Yang, NH., Yuan, YL., Lu, Y. , Zhu, XY. & Wang. ZN. (2002). Study on the structure of scientific and technological workers in China based on the age distribution of authors of research papers. *Kexue Tongbao*, 47, 795-800 (in Chinese).

Jin, BH. & Rousseau, R. (2001). An introduction to the barycentre method with an application to China's mean centre of publication. *Libri*, 51, 225-233.

Jin, BH. & Wang, B. (1999). Chinese Science Citation Database: its construction and application. *Scientometrics*. 45, 325-332.

Jin, BH., Zhang, JG., Chen, DQ. & Zhu, XY. (2002). Development of the Chinese Scientometric Indicators (CSI). *Scientometrics,* 54, 145-154.

Kyvik, S. (1990). Age and scientific productivity. Differences between fields of learning. *Higher Education*, 19, 37-55.

Lehman, H.C. (1953). *Age and achievement*. Princeton (NJ), Princeton University Press.

Liang, L., Kretschmer, H. Guo, Y. & DeB. Beaver, D. (2001). Age structures of scientific collaboration in Chinese computer science. *Scientometrics*, 52, 471-486.

Liang, L., Wu, Y., Ding, F. & Wang, Y. (1999). New developments in the study of distribution of major scientific and cultural achievements in terms of age. In: *Proceedings of the seventh conference of the International Society for Scientometrics and Informetrics* (Macías-Chapula, ed.), 286-293. Colima (Mexico): Universidad de Colima.

Melker, A. (1999). Research evaluation in physics and materials science. In: *Science evaluation and its management* (V. Pačes et al. , eds.) 221-226. Amsterdam: IOS Press.

Papoulis, A. (1991). *Probability, random variables, and stochastic processes* (third ed.) McGraw-Hill: Boston (MA).

Rousseau, R. , Jin, BH., Yang, NH. & Liu, XM. (2001). Observations concerning the two- and three-year synchronous impact factor, based on the Chinese Science Citation Index. *Journal of Documentation*, 57, 349-357.

Shiavi, R. (1991). *Introduction to applied statistical signal analysis*. Irwin & Aksen: Boston (MA).

Wu Y., Pan Y., Zhang Y., Ma Z., Pang J., Guo H., Xu B. & Yang, Z. (2003). China Scientific and Technical Papers and Citations (CSTPC): history, impact and outlook. These Proceedings.

Zhao, HZ. (1984). *An introduction to the study of science potential*. Beijing: Science Press, pp.195-222 (in Chinese).

Zusne, L. (1976). Age and achievement in psychology. The harmonic mean as a model. *American Psychologist*, 31, 805-807.

A revised version of this contribution has been published as:

Bihui Jin, Ling Li and Ronald Rousseau (2004). Long-term influences of interventions in the normal development of science: China and the Cultural Revolution. Journal of the American Society for Information Science & Technology, 55(6), 544-550.