Where demographics meets scientometrics: towards a dynamic career analysis

Lin Zhang[1,2] and Wolfgang Glänzel[3,4]

[1]*Lin.Zhang@econ.kuleuven.be*
Centre for R&D Monitoring (ECOOM) and Dept. MSI, K.U. Leuven, Leuven (Belgium)

[2]*linzhang1117@gmail.com*
North China University of Water Conservancy and Electric Power, Dept. Management and Economics, Zhengzhou (China)

[3]*Wolfgang.Glanzel@econ.kuleuven.be*
Centre for R&D Monitoring (ECOOM) and Dept. MSI, K.U. Leuven, Leuven (Belgium)

[4]*glanzw@iif.hu*
Institute for Research Policy Studies (IRPS), Hungarian Academy of Sciences, Budapest (Hungary)

Abstract
In an earlier exercise some demographic methods were reformulated for application in a scientometric context. Age-pyramids based on annual publication output and citation impact was supplemented by the change of the mean age of the publications in the h-core at any time. Although the method was introduced to shed some demographic-scientometric light on the career of individual researchers, the second component, i.e., the age dynamics of the h-core can however be applied to higher levels of aggregation as well. However, the found paradigmatic shapes and patterns do not only characterise individual careers and positions, but are also typical of life cycles and subject-specific peculiarities. In the present study, the proposed approach is used to visualise the careers of scientists active in different fields of the sciences and social sciences and notably the second component, the h-core dynamics, is extended to the analysis of scientific journals from the same fields.

Introduction
Since the h-index has been introduced by Hirsch in 2005, the analysis of scientists' individual careers gained a new impetus. Scientists in different disciplines were very soon ranked according to their h-index. Although such exercises proved hazardous by many reasons (cf. Glänzel, 2005), the presentation of the h-index of scientists, considered 'leading' in their fields, remained quite popular (e.g., Glänzel & Persson, 2005; Egghe, 2006; Bar-Ilan, 2006, 2010; Cronin & Meho, 2006, 2007; Levitt & Thelwall, 2009 for the field of information science and scientometrics). However, high h-index values might be considered kind of confirmation of the supposed prominent position the researcher holds in the community. Besides its well-known subject-dependence (cf. Batista et al., 2006), the h-index is sensitive to the scientist's academic age as well. In order to compensate for this effect, it was suggested to normalise the individual's h-index by the respective career length (e.g., Jensen et al., 2009). On the other hand, such normalisation eliminates important aspects that are otherwise captured by the h-index. Subject-specific peculiarities and characteristics of career stages are thus lost if the measure is normalised this way. The changes in the h-index and the h-core along with the age structure of publications and citations allow a deeper insight into an individual's career and might reflect breaks, a caesura or shift in the scientist's academic life.
In a previous note (Glänzel & Zhang, 2010), we have monitored the changes in the age structure of the h-core of individual scientists by introducing a new measure. In addition, we have used an analogon of the well-known age pyramid presentation of the age distribution in a human population, which was re-defined for visualising the change of publication activity and citation impact of individuals over time.

Methods and results

In some recent studies, the career of scientists in the field of information science has been analysed in the light of the dynamics of their productivity and citation impact. In particular, the relationship between creativity and both chronological and professional age in information science has been explored by Cronin & Meho (2007). The authors have used high-impact work and cumulative citation rates to capture the shape of a scientist's career. A more recent study by Levitt & Thelwall (2009), it has been shown that high citation impact of individual papers in information and library science is not always reflected by a high h-index of their first authors.

In the present study we will apply the tools developed in the previous study (Glänzel & Zhang, 2010) to two levels of aggregation. The proposed approach is used to visualise the careers of (1) individual scientists active in different fields of the sciences and social sciences, and is extended to the analysis of (2) scientific journals from the same fields. We will analyse 'career-specific' patterns of publication activity and citation impact along the following research questions.

- In how far does citation impact mirror the evolution of publication activity?
- Is there a phase shift between publication and reception of results?
- Is there any cumulative effect of impact independently of publication activity?
- In how far is the influence of 'hot papers' or 'hot topics' measurable by the age distribution and the changing age of the h-core?

In order to answer these questions we will apply two approaches. The first approach is based on a well-known visualisation tool in demographics. It is designed to measure and to compare publication and citation life-time for a selected individual. Before we apply this tool, we shortly recall the basic patterns according to typical population pyramids obtained in demographics. The second tool is based on the calculation of the arithmetic mean age of publications of the *h-core sequence* which will be explained later.

The Age Pyramid

In demographics, the population pyramid is an elementary tool to reflect the age structure and the growth characteristics of a given population. In an age pyramid or age structure diagram, the age distribution in a human population is shown in a double bar diagram, where the various male age groups are plotted against the corresponding female groups. Demographers distinguish about 5–7 paradigmatic shapes reflecting different types of expanding, stationary and contracting population models. From the mathematical viewpoint, one can distinguish simple linear, convex and concave shapes as well as more complex shapes with and without inflection point. Most known shapes in demographic analysis of human populations are the triangle (reflecting steady growth with high fertility and high mortality in all age groups), the pagoda shape (with very high fertility and high infant mortality), the bell shape (typical of the baby boom in the industrial countries after World War II), beehive shape (reflects a stationary structure, provided infant mortality is low) and the "onion" shaped (reflecting superannuation of the population).

Here we have to stop for moment since the above characteristics refer to 'real' populations. The adoption of demographic model in an informetric context requires some re-interpretations. While the notion of fertility can still be interpreted as the current publication activity, *mortality does not exist* in this context since papers, once published, and citations, once received, will not disappear from the system any more. And life expectation can at the best be interpreted in terms of obsolescence as reflected, for instance, by the life-time distribution of citations (cf. Glänzel & Schoepflin, 1995). The concept of mortality should therefore be renounced when this model is applied to informetrics. This holds, above all, for the interpretation of the triangle and pagoda shape. Both patterns just express higher activity

in recent years; by contrast, the beehive shape reflects stable publication activity over time, while the onion or urn shape indicates decreasing activity of the author in the recent years. The age profile of *citation impact* is subject to a very special effect. Papers might still be cited even when a scientist is not active any more, the right-hand side of the diagram tends to be rather triangle- or pagoda-shaped but beehive or onion shape might occur as well. Furthermore, citation impact is expected to have a non-degenerate age distribution when an author has become inactive and publication activity is flatlined.

At this point we have to stress that one has to distinguish between the case of individuals with their specific life cycles and higher aggregates like journals, topics or institutes, where individual life cycles do not play an important part and 'terminators' (cf. Price & Gürsey, 1976) are continuously replaced by new authors and, therefore, both activity and impact distributions are not mainly determined by the individual's life cycle.

For the following exercise we use anonymous samples, where we have selected four individual authors from Thomson Reuters' Web of Science representing a group of scientists with about 25 or more years of professional experience in four different subject areas, one each representing the life sciences, natural sciences, mathematics (statistics & probability) and the social sciences. The selected authors are, on one hand, *not representative* for their field in the sense that they would represent the standard of their field; they have a quite long career with large publication output and higher than average citation impact. On the other hand, they *are indeed representative* for their field in the sense that their communication behaviour, i.e., their publication and citation behaviour reflects the subject-specific peculiarities of their discipline.

The pyramids are presented in Figure 1. In order to facilitate visualisation and interpretation, we have grouped publication and citation counts by three-year periods. This helps avoid fluctuations and avoids the occurrence of periods of relative inactivity as well. We plot the distributions of papers according to their age at the left-hand side of the diagram, that of citations at the right-hand side. Furthermore, we have rescaled citations by factor 25.

Some typical patterns can immediately be recognised in Figure 1. The first peculiarity, that strikes the eye, concerns the asymmetry of publication and citation patterns. While the publication-age distribution of the first three authors is of beehive/onion type, the fourth author represents a triangle type. While the second and third author are already active for 35-45 years and their recently decreasing activity seems to be plausible, the onion shape of the first author is somewhat surprising. The triangle shape of the fourth author is not an exception to the rule. Besides the individual peculiarities also the influence of the subject field is visible. The shape of the distribution in the natural sciences and in mathematics is more stretched and the absolute frequencies are lower than in the life sciences and the social sciences. This effect is even more obvious if the age of citations is considered. The gap between the citation impact in the natural sciences, notably in mathematics, on one hand, and the life sciences, on the other hand, is large. The triangle and pagoda type can be found in the first and the fourth diagram, respectively. While the beehive in the second case and the onion in the third one by and large mirror the corresponding shapes of publication age, the onion shape of publication age is contrasted by citation triangle in the first case. This might have two reasons, particularly a phase shift between publication and reception of results, or some recent hot topics in the work of the author in question. The latter case can be observed for the fourth author in the social sciences. Also the author in the natural sciences is characterised by contrary trends of publication and citation age in the most recent periods. The trend might point to a certain citation delay in this case.

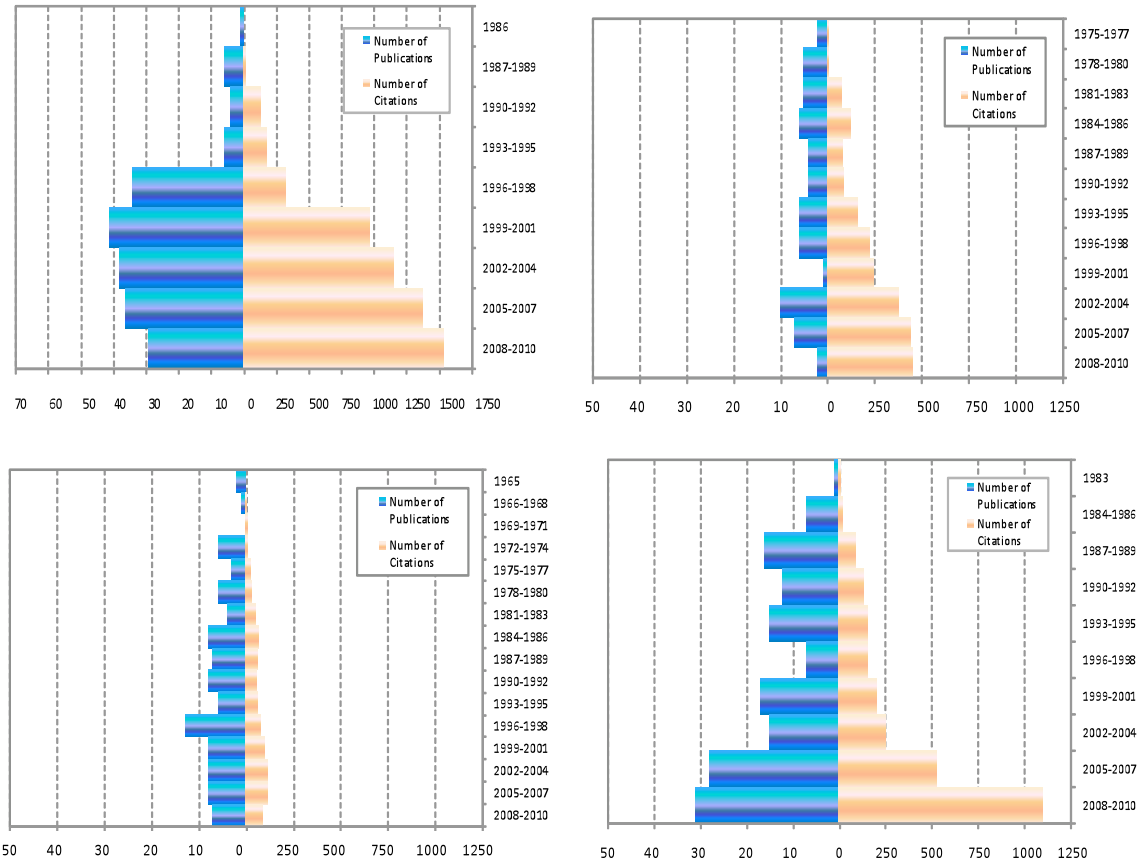

Figure 1. Scientometrics age pyramid for four scientists (top left: life sciences, top right: natural sciences, bottom left: mathematics, bottom right: social sciences)
[Data sourced from Thomson Reuters Web of Knowledge (formerly referred to as ISI Web of Science)]

Further examples of triangle/pagoda or beehive/onion shapes of the scientometric age pyramid for individual authors in the field of information science can be found in the piece by Glänzel and Zhang (2010).

The age pyramid of scientific journals will structurally differ from the previous one as has pointed to above. The distributions of publication and citation age will not reflect peculiarities of individual life cycles. However, the same basic patterns can be observed for journals too. In order to illustrate this we have chosen four journals from the same domains, namely *Journal of Infection* for the life sciences, *Superconductor Science & Technology* for the natural sciences, *Probability Theory and Related Fields* for mathematics (statistics & probability) and *Scientometrics* for the social sciences. The results are presented in Figure 2. *Journal of Infection* is dynamically growing journal with linear shapes of the age distribution of both publications and citations. The evolution of Scientometrics is even more dynamic; both distributions have pagoda shape. This is certainly a result of the sharp rise of quantitative science and technology studies having taken during the last three decades. The interpretation of the pyramid of the journal Superconductor Science & Technology is not so easy. The frequency of younger citations increases linearly but the publication age has rather a beehive, almost an onion shape. The reasons for that are not clear. The publication age distribution of *Probability Theory and Related Fields* reflects superannuation; however the shape of the citation age is linear. This substantiates that increasingly recent citations are predominant. There is indeed cumulative effect of impact independently of publication activity, and citation impact may not mirror the evolution of publication activity in the same period of career. On

the other hand, the linearly increasing impact also shows the strong position the journal holds among the theoretical journals in the field of statistics and probability.

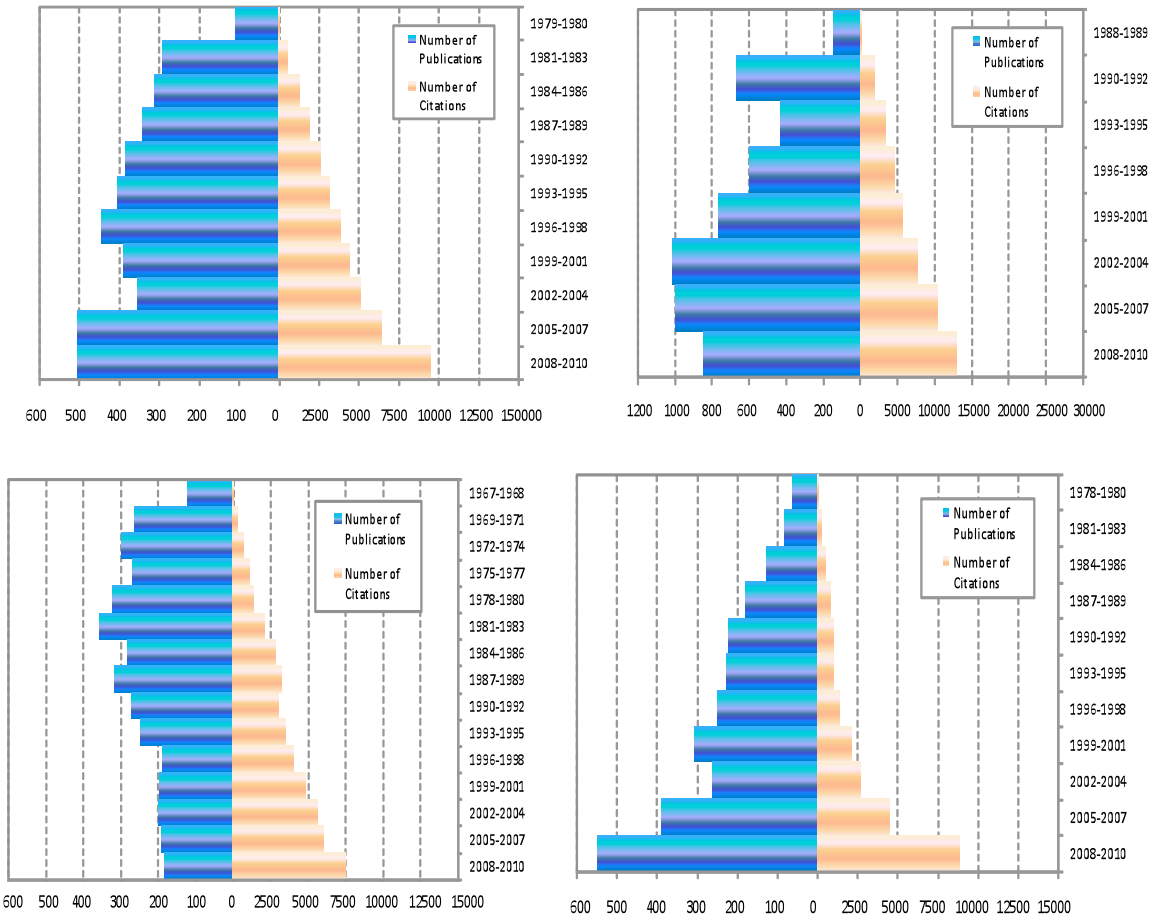

Figure 2. Scientometrics age pyramid for four journals (top left: Journal of Infection, top right: Superconductor Science & Technology, bottom left: Probability Theory and Related Fields, bottom right: Scientometrics)
[Data sourced from Thomson Reuters Web of Knowledge (formerly referred to as ISI Web of Science)]

The average age of publications in each Hirsch core

The Hirsch core, or *h*-core, is formed by those papers that have received at least *h* citations, where *h* denote the actual value of an h-index (Jin et al.,2007). Liang (2006) proposed the h-index sequence in order to measure the dynamics of the h-index in a scientists career. She defined the h-index sequence h_k as the h-index of the papers published by the author in question in the time interval $[n\text{-}k+1, n]$, where n is the most recent year. Citations are counted for the same period. However, unlike Liang's approach aiming at constructing the h-sequences in a retrospective fashion, we plotted the evolution of an author's or a journal's h-index through his/her or its career by calculating the h-index from the beginning year of their career, that is, in a prospective manner. In other words we first calculated the h-index for papers published in the first year of their career, then the first two years, the first three years, and so on until the most recent year is reached. This can be done even if the author's productive career has come to an end before. This method of calculation is also in line with Burrell's study (Burrell, 2007) of time-dependence of the h-index using stochastic models.
Then a h-core sequence is defined analogously to the h-core for the h-index sequence. Here we include the most recent ones in the h-core if there are several publications with exactly *h* citations. The calculation of the age of publications in each h-core is also based on the three-

year sub-periods. The age of papers in each h-core is equivalent to the difference between the "current unit" (i.e., the period for the h-core in question) and the time unit of publication. For instance, a publication in unit 1 has an age of 2 if it appears in the h-core of unit 3. Here we define the "time zero" for the author in question as the time unit when the author's first publication appeared in the WoS.

The arithmetic mean age of publications of this *h-core sequence* is calculated, which expresses whether the more recent or the older publications are predominant in the respective h-core. The obtained patterns reflect different aspects of changing impact from the pyramid approach. Also for this indicator, we can find four paradigmatic patterns.

- A linear shape of the mean age of the h-cores plotted against time reflects steady growth of the age of most cited publications.
- A convex shape reflects accelerated growing age of the most cited papers. This means that the "top" papers were rather published in earlier stages of the scientist's career.
- A concave shape reflects decreasing age of highly cited papers, that is, recent papers by the author are the more cited ones.
- "Indefinite" shape. This covers all cases not listed above.

Case 1 can be considered a standard situation. It can be expected that a paper remains in the h-core once it has already received a sufficient number of citations. If an author becomes less active or inactive, the age of the h-core will disproportionally increase after a while. In extreme cases this might result in a convex shape. If, however, an increasing number of *recent* papers enter the h-core, the age curve turns concave. For instance, a new emerging topic or a 'hot paper' and its follow-ups might cause this phenomenon (cf. Glänzel and Zhang, 2010).

The mean age sequences of the h-core for the four selected authors are presented in Figure 3. The most striking common feature is probably the subject dependence of the mean age. The mean age of the h-core for the author in the life sciences does even not exceed 4 years in the most recent year. By contrast, the mean age in the natural sciences and mathematics with 7–8 years in 2010 is pronouncedly higher. However, this might also partially be caused by the different career lengths of these scientists, where the scientist in life science has a relatively shorter career compared to the scientists in the natural sciences and mathematics. The fourth case is most interesting because of the shape. While the shapes in the natural sciences and mathematics are nearly linear, this one turns from a linear to a concave graph. The reason is a 'hot topic' on which the author published in the second half in the last decade. The first case is also interesting as it basically experienced a "concave shape" around the middle phase of the career and then turned to a linear graph. When taking a deeper look into the "concave" phase of this author, we found relatively many highly cited publications appearing during the period 1996-2001. For instance, 5 new papers entering the h-core in 1998 (with 13 publications in total) were published in the same period 1996-1998 and it was striking that in 2001, among the 21 publications in the h-core, 15 papers have been published in 1996-2001. This give an explanation of the temporary concave graph of the h-core age sequence. However, there have been much less "recent" new-comers into the h-cores afterwards, thus a linear graph is observed since 2001. It is also interesting that most of the highly-cited publications appearing in the "concave" phase (1996-2001) have always remained in the subsequent h-cores, and several of them are among the most cited publications of the author. On the other hand, there are still new members from the "concave" period entering the most recent h-cores. By 2010, 26 papers from 1996-2001 exist among the 40 publications in h-core, which means roughly two thirds of the most important publications of the author were from the "concave" period. It could be concluded that the author had published in "hot topics" during the corresponding period and apparently these topics remained "hot" or important even after a decade. Combined with the age pyramid analysis, the most productive phase of the author was found in the period 1999-2001. Though the author is still active in his field, we

may conclude that the period of "concavity" might be the one with highest impact in his publication career. To confirm this conjecture, we further have a look at the h-index sequence of the author (Figure 3a), not surprisingly, there was indeed an accelerated increasing trend of h-index in the period of 1996-2001.The concave turn of the mean age graph of the forth author is even more striking. To make a comparison of the career evolution of these two authors, the h-index sequence of the latter author is also presented in Figure 3a. The h-index is accelerated increasing in the most recent periods, corresponding to the "concave" phase of his h-core age, and on the other hand, is also in line with his most productive and most cited period. Back to our research question, the influence of "hot papers" or "hot topics" are indeed measurable by the age distribution and the change of the age of the h-core in time and the "hot papers" also mark some important points in the author's career.

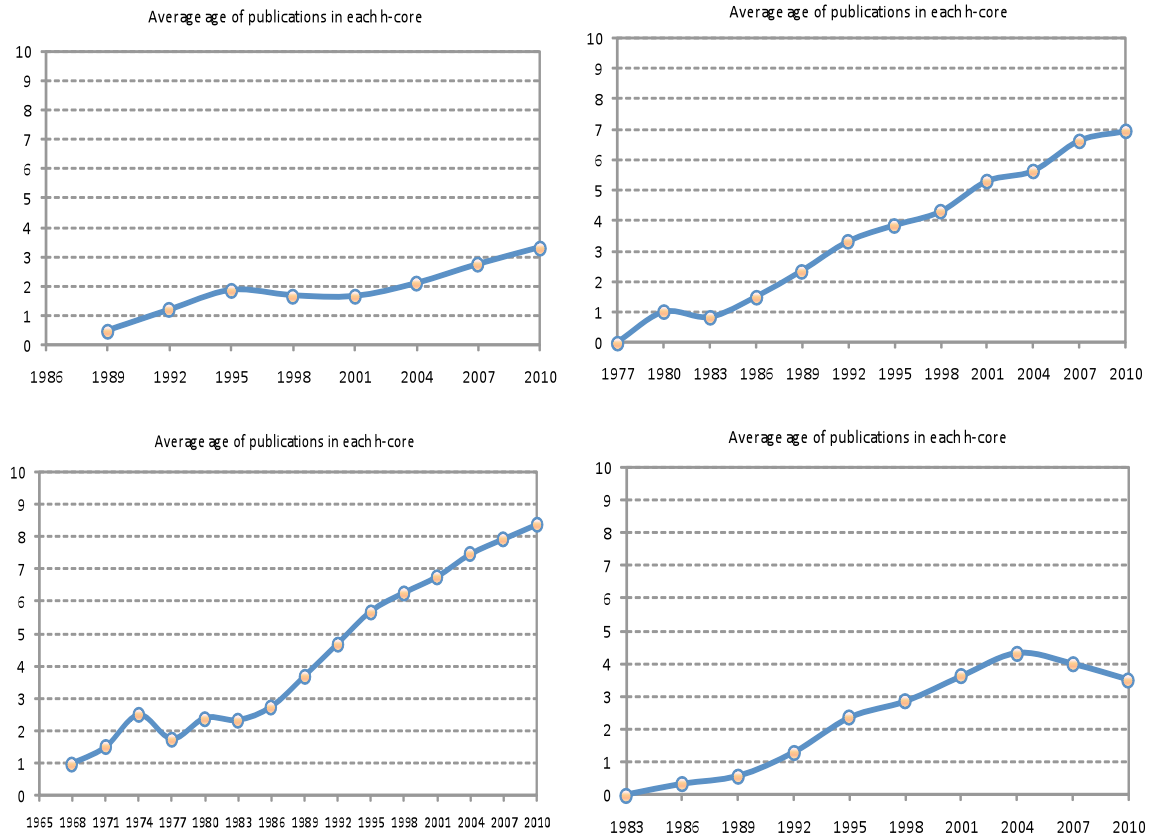

Figure 3. Mean age sequence of the h-core for four scientists (top left: life sciences, top right: natural sciences, bottom left: mathematics, bottom right: social sciences)
[Data sourced from Thomson Reuters Web of Knowledge (formerly referred to as ISI Web of Science)]

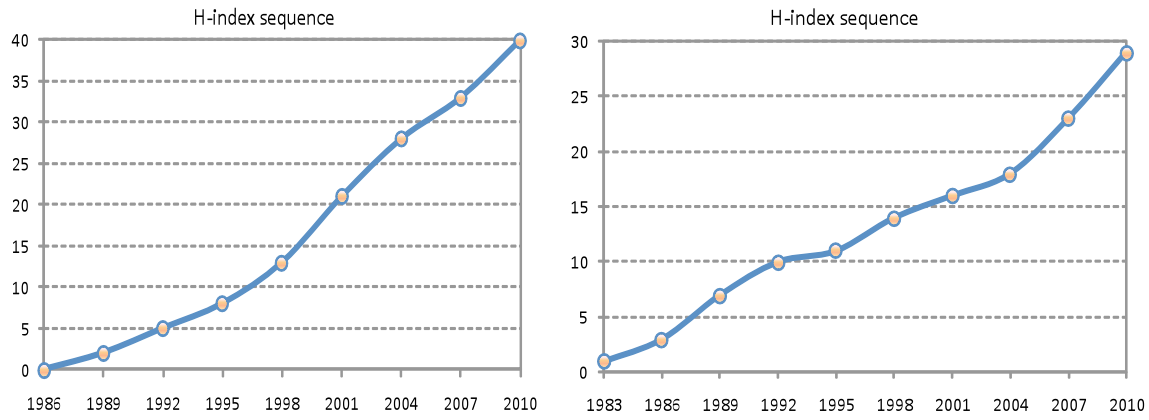

Figure 3a. H-index sequence for scientists in life sciences (left) and in social sciences (right)
[Data sourced from Thomson Reuters Web of Knowledge (formerly referred to as ISI Web of Science)]

The idea of using h-indexes for scientific journals was introduced as early as in 2005 (cf. Braun et al. 2005, 2006). Although it was shown that there is a strong correlation between the h-index and hybrid measures calculated on the basis of the number of publications in a journal and their citation impact (Schubert & Glänzel, 2007), through its "size dependence" the journal h-index reflects somewhat different aspects than other journal impact measures, notably the ISI Impact Factor. And these aspects are indeed closer to what we can consider as or characterise by the expressions "career", "demography" or "life time".

The mean age sequence of the h-core for the four journals can be found in Figure 4. There are, of course, similarities between the charts for the individuals and the journals; above all, the "high age" in mathematics. At a first sight, the low mean age of the h-core for the physics journal is somewhat striking. However, in the light of the scope of the journal the reasons for the almost concave curve with low gradient become somewhat clearer. The journal is (1) focussed on a hot subject, (2) primarily publishes an experimental papers and accepts theoretical articles only if those are clearly linked to experiments, and (3) publishes *rapid communications* along with regular research articles. The fast ageing is also reflected by the low Cited Half-life reported for this journal in the annual *Journal Citation Reports* of Thomson Reuters. In the ISI Subject Category *physics, condensed matter* this journal, together with journals on nanoscience and –technology, has one of the lowest Cited Half-life.

The shape of the age diagram of the h-core sequence of the fourth journal, *Scientometrics*, resembles to that of the fourth author in Figure 3. For this journal we have found several hot topics that might be responsible for the dramatic drop around 2004. Above all, the *h-index and its derivatives* has speeded up scholarly communication and provoked extended discussions with numerous citations. In addition, highly topical issues like *science and technology in emerging economies, author self-citations* and *mapping and visualisation of networks* have also contributed to this effect. Actually, if we calculated the average age of new papers entering in each h-core, the mean age of these "new-comers" in h-core of 2007 was only 2.85, and that same value in 2010 was even lower, particularly, 2.25, in contrast to the much higher mean age (4.4) of the new members of the h-core in 2004. Due to a certain citation delay, some "hot papers" appearing around 2004 have experienced a citation boom in the subsequent periods. In particular, one third of the highest-cited publications in 2010 were from 2002-2006, where the two top cited papers were published in 2004 (Ho, Citation review of Lagergren kinetic rate equation on adsorption reactions) and in 2006 (Egghe, Theory and practise of the g-index). In this context we would like to mention that among the 55 publications in the h-core of 2010, the year 2006 was, with 6 highly cited contributions, the most important contributor. Another interesting phenomenon is that though the last decade was a prosperous period for *Scientometrics*, there is no paper published in the year of 2000 appearing in h-core in any period under study. The graph of h-index sequence of *Scientometrics* is presented in Figure 4a, where we can observe a clear turning point of increasing h-index since 2004. For both the scientist and the journal in social science, the "concave" period of their h-core age sequence is exactly corresponding to their phase with the highest citation impact.

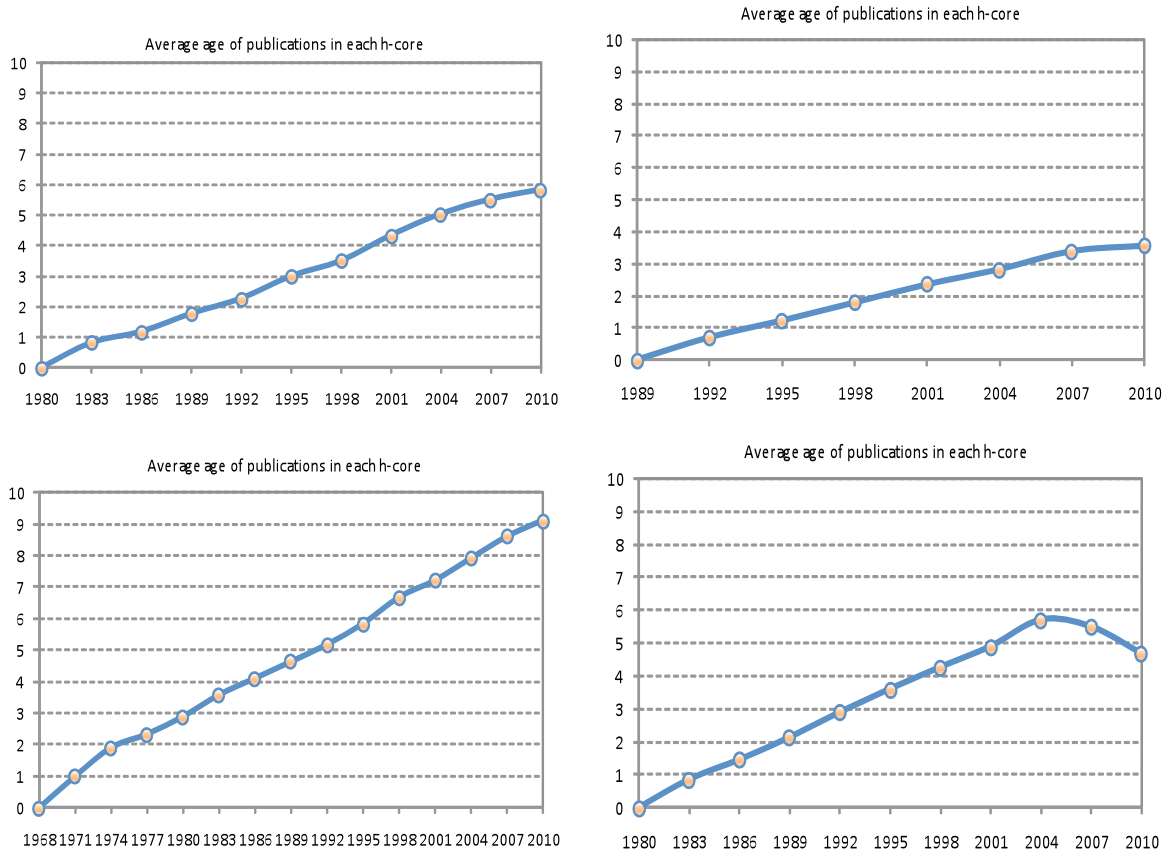

Figure 4. Mean age sequence of the h-core for four journals (top left: Journal of Infection, top right: Superconductor Science & Technology, bottom left: Probability Theory and Related Fields, bottom right: Scientometrics)
[Data source: Thomson Reuters, Web of Knowledge]

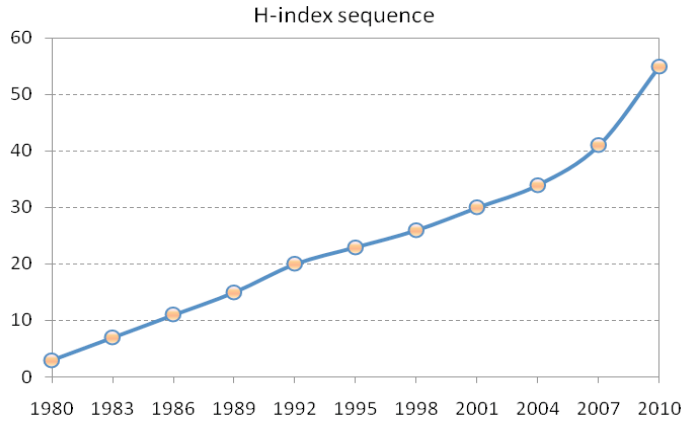

Figure 4a. H-index sequence for Scientometrics
[Data source: Thomson Reuters, Web of Knowledge]

Conclusions

Both the scientometric age pyramid and the age curve of the h-core proved interesting tools for the dynamical career analysis of individuals and journals. Another possible extension of this type of demographic–informetric analysis might form the application of 'real' author populations such as research groups, departments or even institutes. The h-index and its derivatives have already been introduced in institutional evaluation (e.g., Molinari & Molinari, 2008). Thus the applicability of h-sequences and the mean age of their h-cores to

institutions will work as well as in the case of journals. Perhaps more pronounced shapes and trends can be expected. The changing constitutions of teams with stable cores (continuants) but with varying extend of transients, newcomers and terminators (cf. Price & Gürsey, 1976) might have strong influence on the age structure of the team's publications and citations received by those. It might then be interesting to identify young, dynamic teams focussing on hot research and to monitor the ups and downs in activity and impact of groups and departments with long tradition and experience in their research areas.

References

Bar-Ilan, J. (2006). H-index for Price Medalists revisited. *ISSI Newsletter*, 2(1), 3–5.

Bar-Ilan, J. (2010). A follow-up on the h-index of Price medalists. *ISSI Newsletter*, 22 (2), 39–43.

Batista PD., Campiteli MG. & Kinouchi O. et al. (2006). Is it possible to compare researchers with different scientific interests? *Scientometrics*, 68 (1), 179-189.

Braun, T., Glänzel, W. & Schubert, A. (2005), A Hirsch-type index for journals. *The Scientist*, 19 (22) 8.

Braun, T., Glänzel, W. & Schubert, A. (2006), A Hirsch-type index for journals. *Scientometrics*, 69 (1), 169–173.

Burrell QL.(2007). Hirsch index or Hirsch rate? Some thoughts arising from Liang's data. *Scientometrics*, 73 (1), 19–28.

Cronin, B., & Meho, L. (2006), Using the h-index to rank influential information scientists. *Journal of the American Society for Information Science and Technology*, 57 (9), 1275–1278.

Cronin, B. & Meho, L. (2007). Timelines of creativity: A study of intellectual innovators in information science. *Journal of the American Society for Information Science and Technology*, 58 (13), 1948–1959.

Egghe, L. (2006). Theory and practise of the g-index. *Scientometrics*, 69 (1), 131-152.

Glänzel, W. & Schoepflin, U. (1994). A stochastic model for the ageing analyses of scientific literature. *Scientometrics*, 30 (1), 49–64.

Glänzel, W. & Persson, O. (2005). H-index for Price medalists. *ISSI Newsletter*, 1 (4), 15-18.

Glänzel, W. (2006). On the opportunities and limitations of the h-index (in Chinese). *Science Focus*, 1 (1), 10–11.

Schubert, A. & Glänzel (2007), W., A systematic analysis of Hirsch-type indices for journals. *Journal of Informetrics*, 1 (3), 179–184.

Glänzel, W. & Zhang, L. (2010). A demographic look at scientometric characteristics of a scientist's carrier. *ISSI Newsletter*, 2010, 6 (3), 66–84.

Ho, YS. (2004). Citation review of Lagergren kinetic rate equation on adsorption reactions. *Scientometrics*, 2004, 59 (1), 171-177.

Jensen, P., Rouquier, J.B. & Croissant, Y. (2009). Testing bibliometric indicators by their prediction of scientists promotions. *Scientometrics*, 78 (3), 467–479.

Jin BH, Liang LM, Rousseau R, et al. (2007). The R- and AR-indices: Complementing the h-index. *Chinese Science Bulletin*, 52 (6), 855–863.

Levitt, J.M. & Thelwall, M. (2009). The most highly cited Library and Information Science articles: Interdisciplinarity, first authors and citation patterns. *Scientometrics*, 78 (1), 45–67.

Liang, L. (2006). H-index sequence and h-index matrix: Constructions and applications. *Scientometrics*, 69 (1), 153–159.

Molinari, A. & Molinari, J.F. (2008). Mathematical aspects of a new criterion for ranking scientific institutions based on the h-index. *Scientometrics*, 75 (2), 339–356.

Price, D.D. & Gürsey, S. (1976). Studies in scientometrics. Part 1. Transience and continuance in scientific authorship. *Internation Forum on Information and Documentation*, 1, 17–24.