# Referencing patterns among individual researchers: study of the use of scientific literature from a micro-level perspective

Rodrigo Costas[1] and Thed N. van Leeuwen[2] and Maria Bordons[3]

[1] *rcostas@cwts.leidenuniv.nl,* [2] *leeuwen@cwts.leidenuniv.nl*
Leiden University, Centre for Science and Technology Studies (CWTS), Wassenaarseweg 62A, 2333AL Leiden (The Netherlands)

[3] *maria.bordons@cchs.csic.es*
Instituto de Estudios Documentales sobre Ciencia y Tecnologia (IEDCYT), Center of Human and Social Sciences (CCHS), CSIC, Albasanz 26-28, 28037 Madrid (Spain)

## Abstract
Bibliometric methodologies are useful to study how researchers cite and include bibliographic references in their publications. This study seeks to analyze the different patterns of use of references by individual researchers, focusing on the number and type of references that they include in their papers and relating these patterns with other individual factors such as the age and the scientific performance of researchers. The analysis comprises of 1,064 researchers with a permanent position at the Spanish CSIC in three different scientific areas. Their scientific publications in the Web of Science database during an eleven-year period are collected and different reference-based indicators are obtained. Results show that top-performing scientists use in their papers a broader range of scientific literature as compared to the other researchers, suggesting that a good knowledge on the current literature in a field is necessary in order to enhance the scientific performance of scientists. On the other hand, veteran researchers tend to rely more on older literature and on literature that is not covered by the Web of Science.

## Introduction
Bibliometric indicators are useful tools for a better understanding of the scientific process and in particular for the study of the different strategies followed by researchers when communicating new knowledge and discoveries. Different aspects such as the determinants of scientific performance (Gonzalez-Bambrila & Veloso, 2007; Long et al, 2009), how scientists interact among them (Calero et al., 2006; Zuccala, 2006; Jiang, 2008) and how their roles in the production of new results change as they grow old (Levin & Stephan, 1989; Gingras et al, 2008; Costas et al, 2010) can be studied through bibliometric methodologies.
The study of the referencing patterns of scientists can provide us with useful information about the communicative practices existing in a given discipline (Velho & Krige, 1984) as well as allow us to explore differences between scientists even within a discipline. Among the research topics of high current concern in the literature we can mention the typology of cited references (Clements & Wang, 2003; Larivière et al, 2006), scientists´ ways of searching and using references (Shanmugam, 2009; Budd & Magnuson, 2010), reasons and attitudes (Oppenheim & Smith, 2001, Clarke & Oppenheim, 2006), limitations and bad uses of referencing practices (Roth & Cole, 2010) and the relationship between referencing patterns and impact of papers (Alimohammedi & Sajjadi, 2009; Corbyn, 2010).
Through bibliometric methodologies it is possible to analyze the number and type of references that are given in the papers of scientists in order to understand how they document their publications. The reference lists of scientific publications represent a good example on how scientists are using scientific information, what are their influences (Budd & Magnuson, 2010) and also (up to some extent) inform us about the knowledge that they have about their respective fields of work. According to Moed (2006) a reference list marks the 'socio-cognitive location' of a paper. An extension of this idea is that the reference lists of the papers of an author signal his/her 'socio-cognitive' environment. It is important to study the

referencing behaviour of researchers as it is an important part of the total picture of the scholarly communication which is still not totally understood. Assuming that a high number of references per paper can be associated to more comprehensive papers and good knowledge of the field, we want to test the hypothesis that the "best" researchers in a scientific area use more references in their papers as compared to their colleagues with a lower scientific performance.

**Objectives**

The main objectives of this study are the following: a) to analyze the use of information (use of references) by individual researchers in three different areas, focusing on the number and type of references that they include in their papers; and b) to explore whether different factors such as age or scientific performance of scientists might be related to their referencing behaviour. Different research questions are addressed: What type of scientific literature use researchers according to their research area? Are there differences in the literature use according to the scientific performance and age of scientists? Do top scientists use more references in their papers than the rest of scientists in the same research area?
The answers to these questions will provide important insight into the referencing behaviour of researchers, useful for policy makers and research managers as well as for library policies, editors of scientific journals and scientists themselves.

**Methodology**

This study analyses the scientific publications of 1,064 researchers employed with a permanent position (as "civil servants") at the Spanish CSIC in 2005. These researchers are organized in three main scientific fields: Biology & Biomedicine (388), Natural Resources (348) and Materials Science (327) and belong to three professional categories: Tenured Scientist (the most basic category – 558 researchers), Research Scientist (the intermediate category – 269) and Research Professor (the highest category – 237).
For each researcher the scientific production published in journals processed by the Web of Science (WoS) during the period 1994-2004 was downloaded (several methodologies for the proper matching of authors and documents were considered - Costas & Bordons, 2006). Documents published by scientists at CSIC, but also publications with a foreign address as a result of a research stay abroad were both considered. For each scientist a bibliometric profile is built following the methodology for the analysis and research assessment of individuals described by Costas et al. (2010). The bibliometric profile comprises of nine indicators which are grouped into three different dimensions: production, observed impact and expected (journal quality) impact. According to the above mentioned methodology, scientists are classified in three scientific performance classes: Top, Medium and Low. Top researchers are the ones with a high performance in at least two of the three dimensions, Medium class present an intermediate performance in two of the three dimensions and Low class researchers have a low performance in at least two of the three dimensions suggested (cf. Costas et al., 2010). The main strength of this methodology is that it offers a balanced and complete view over the research performance of individual scientists as three different dimensions are considered for the analysis. Moreover, a classificatory scheme of researchers is provided instead of the wide-spread rankings of scientists in which differences between relative positions are very often no significant. Thus, some of the most common problems, limitations and side effects related with the use of bibliometric indicators (Weingart, 2005) are minimized.
Regarding the analysis of the cited references by the individual scientists, a window of 10 years has been considered for the calculation of the references cited by the researchers. This window is set considering the year of publication of the source papers and goes 10 years

backwards in the cited references. Thus for papers published in 1994 only cited references between 1994-1984 are considered, 1995-1985 for papers in 1995, 1996-1986 for papers in 1996, and so on. With this reference window we want to minimize the effects of the different ages of the researchers (more veteran researchers could use older literature simply because they published in the earlier years of the period and so). In this regards it is also important to mention that almost all researchers under analysis (91%) had publications already in the years 1994-1995, so we can assume a quite homogeneous population in terms of age-production during the whole period of analysis (1994-2004).

Considering these criteria, the following indicators based on the analysis of the cited references have been calculated:

- References per document: number of all references included in all source documents of each researcher divided by his/her total number of publications (notes[16], letters, reviews and articles included).
- References per article: mean number of references per document type Article.
- External references per document: external references are references to documents that were not published by any of the co-authors of the source document (for this indicator only Web of Science documents included in the references were considered and only for the period 1994-2004).
- Total unique references: total number of unique references cited by every researcher.
- Total unique references per document: total number of unique references cited by the researcher divided by his/her total number of publications.
- Average year of the cited references: this is the mean of the ordinal publication year of the references (using a window of 10 years for each paper). Thus, references from the same year of publication as the paper are considered as 0, from a year before as 1 and so on.
- Percentage of references to non-WoS literature: percentage of references to documents not included as source documents in the Web of Science (i.e. books, non-WoS journals, reports, etc).

For the analysis of age, researchers were considered in the following three groups (according to their age in 2004).

- Young: researchers with ages between 32 and 43.
- Senior: researchers between 44 and 56 years old.
- Veteran: researchers with ages between 57 and 69.

These thresholds correspond to the distribution by percentiles of researchers by age (P25 and P75).

**Results**

First of all we present the main figures about the production of the researchers under analysis that can help to contextualize these results. The researchers of the three areas account for a total of 24,982 documents: 9,660 in Materials Science, 9,318 in Biology & Biomedicine and 6,102 in Natural Resources; receiving 80,546, 189,699 and 56,940 total citations respectively. To see additional results at the individual level we refer to Costas et al (2010). The distribution of these documents by 'document type' shows that articles are the predominant type (91% of all documents), followed by meeting-abstracts, reviews and other types (3% each of them)[17].

[16] Please note that the document type Note was removed form the WoS in the mid 1990's, from that moment most Notes are classified as Normal articles. However, they are present in the early years of the dataset used in the study.

[17] By areas, the distribution is 83% Articles, 8% Meeting Abstracts, 6% Reviews and 3% rest in Biology & Biomedicine. 96% Articles, 1% Meeting Abstracts, 1% Reviews, 2% other types in Materials Science; and 94% Articles, 1% Meeting Abstracts, 2% Reviews and 3% other types in Natural Resources.

*1. How is the distribution of cited references by research areas?*

In our approach, the references included in the documents published by the researchers are considered as an indication of the use of information and as a proxy of the knowledge that they have on their research fields. The distribution of the total average number of references per document by researcher and research area is presented in Figure 1.

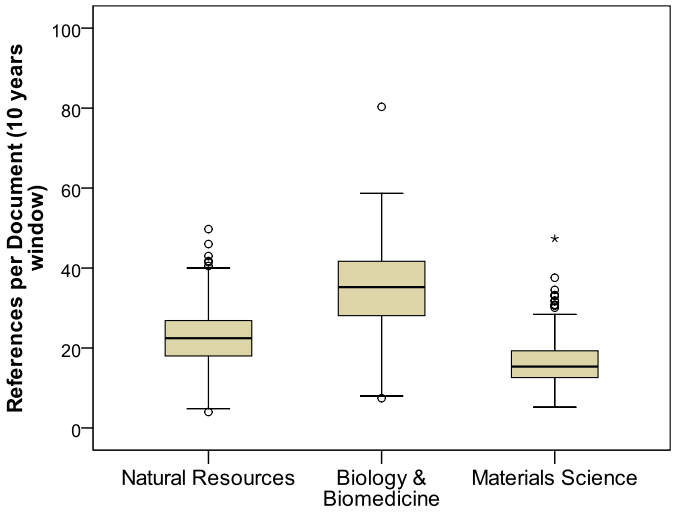


**Figure 1. Distribution of references per document by area.**

Biology & Biomedicine is the area where researchers have the highest average number of references per document, followed by Natural Resources and Materials Science. Statistical significant differences have been found among the researchers of the three areas (Mann-Whitney U test, $p< 0.05$). The previous analysis clearly shows that the number of references per document included by a researcher in his/her publications is clearly field-related, and therefore we can say that an element that determines the use of references is the field in which the researchers are working.

*2. Do Top researchers use more references in their publications as compared to the other scientific performance classes? And do they use a broader range of references?*

As shown in the previous analysis, the number of references per document of individual researchers is linked to the research area. Now we study whether top researchers use more references than their peers within the same research area or have relatively the same number of references as their colleagues in the same fields. Figures 2a to 2e present the distributions of the rates of cited references per paper of individual researchers by areas and also considering their classification in scientific performance classes.

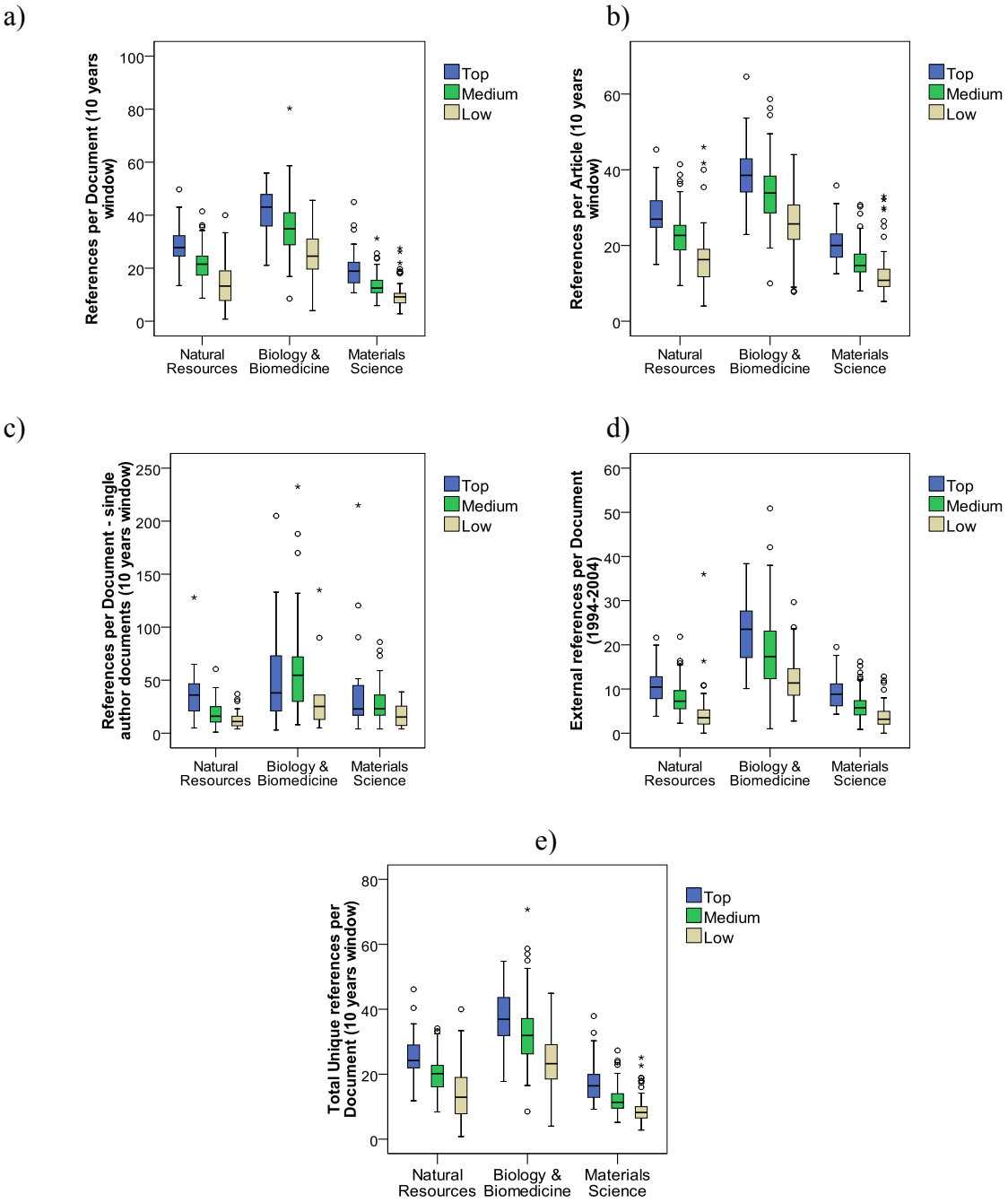


**Figure 2(a-e). Reference analysis by scientific performance class and area.**

According to Figure 2a, Top researchers clearly tend to present more references per document than the other two scientific performance classes in the three areas. These differences are statistically significant in all cases (Mann-Whithney U, $p<0.000$).

The same analysis has been performed including only the document type "Article" in order to avoid the possible bias towards other document types like "Reviews" that normally include more references than regular articles (Figure 2b). Also in this case Top researchers present the highest number of references per article ($p<0.000$), thus proving that Top researchers consistently use more references in their articles than Medium and Low class scientists.

In order to control for the possible influences produced by the collaboration of researchers (e.g. researchers with a high degree of collaboration could include more references in their papers as they are suggested by their co-authors – although in this case we could also argue

that these references enter as a part of the know-how of all the co-authors of the paper), an analysis based only on the single-authored papers published by the researchers under study has been performed (Figure 2c). With this analysis we assume that the researchers only use the references and literature that they know by themselves. The problem of this analysis is that as single-authored papers are becoming less frequent in scientific publication, the number of researchers involved in this analysis is lower (only 259 researchers enter in the analysis). In any case, again we can see how Top researchers tend to present more references per document than researchers in the other two classes (statistical significant differences found in Natural Resources, $p<0.05$), the only exception is observed in Biology & Biomedicine where Medium class researchers graphically present more references per document in their single-authored papers (although no statistical significant differences have been found).

In order to avoid the potential effect of self-citations (authors referencing more their own publications) Figure 2d presents the distribution of external references per document, again showing how Top performance scientists present also the highest number of external references in their publications in all cases as compared to the other classes ($p<0.05$).

The total number of distinct or unique references used by each researcher was obtained and normalized by the total number of publications of each researcher. As a result the rate of unique references per document is obtained for every person, and its distribution has been studied considering the three areas and the scientific performance class of the researchers (Figure 2e). It can be observed how again Top researchers present the highest rate of unique references ($p<0.000$ in the three areas), thus clearly suggesting that Top researchers also use a broader range of unique literature in their publications as compared to the other performance classes.

*3. What type of literature do researchers cite regarding their areas, scientific performance class and age?*

In this section the focus is on the one hand on the percentage of references to non-WoS publications (i.e. books, non-WoS journals, theses, scientific reports, etc.); and on the other hand, on the age of the references used by the researchers. Figure 3 presents the distribution of the percentage of references to non-WoS literature in the publications of the researchers.

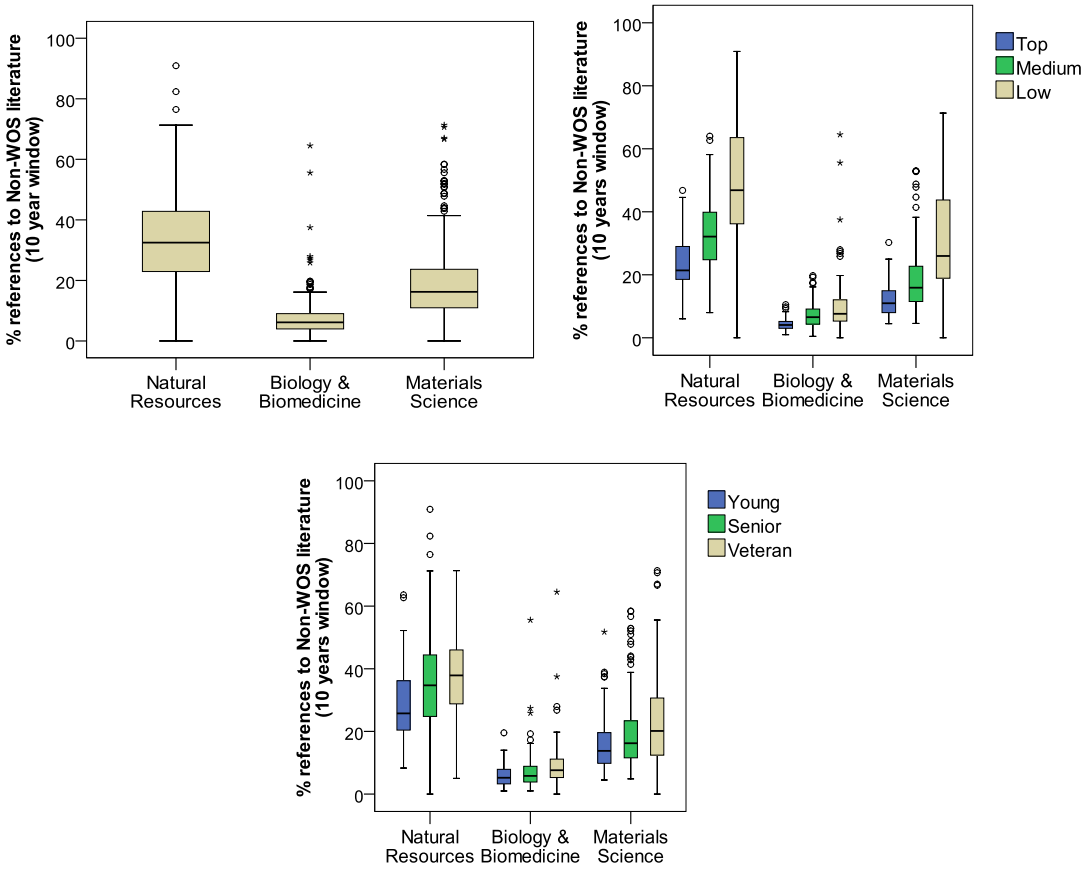


**Figure 3. Distribution of the percentage of references to non-WoS literature by area.**

As the top left graph in Figure 3 shows, differences are found by area with researchers in Natural Resources presenting a higher percentage of references to non-WoS literature, followed by Materials Science and finally Biology & Biomedicine.

Considering the scientific performance classes (top right graph in Figure 3) and the clusters of age of researchers (bottom graph) a very clear pattern can be stated for the three research areas with Top (significant differences in all the cases, $p<0.05$) and younger researchers (also significant differences in almost all the cases, $p<0.05$) presenting the lowest percentages of references to non-WoS literature, while the contrary pattern is found for Low-class and veteran researchers.

It is important to remark here that this analysis on non-WoS references has been performed considering the reference window of 10 years, meaning that only references to papers published up to ten years before the year of publication of the source paper are considered. Thus all researchers have the same 'reference window' (i.e. the same number of years for the cited material). However, it could be argued that more veteran researchers have had less articles available (i.e. less WoS material) to cite in their earlier production, as the volume of journals and publications covered by Web of Science has grown during the last years (Wallace et al, 2009). In this sense, journals that in the early stages of our data system were not included in the WoS database could have been included later. This means that for example, a researcher in 1994 could cite a paper in 1985 published in a journal not included in WoS, but included later in 2000, thus giving the benefit to younger researchers who can cite papers from the same journal when they are already covered by WoS. It is difficult to correct this problem but in order to make sure that the differences observed by scientific performance class are not affected by this issue, we have done an analysis (Figure 4) combining data by

age-group and scientific performance class, thus we can assume that researchers with a similar age had reasonably the same amount of available WoS material to cite.

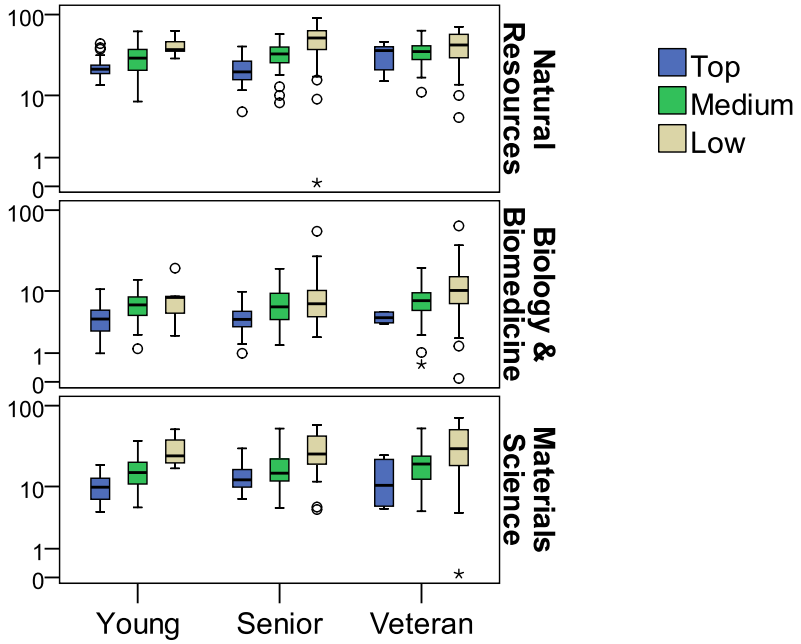


**Figure 4. Distribution of the % of non-WoS references by scientific performance class, group of age and area.**

Figure 4 shows that the same patterns previously described (Low class researchers using more non-WoS references and Top proportionally less) are also observed when controlling for the age. Statistical significant differences have been found in almost all cases when comparing the distribution of Top and Low researchers ($p<0.05$).

The age of the literature cited by the researches has been also studied. In this sense, the average of the ordinal age (within the 10 years reference window) of the cited references of the papers of every researcher has been computed and the distribution by areas is presented in Figure 5.

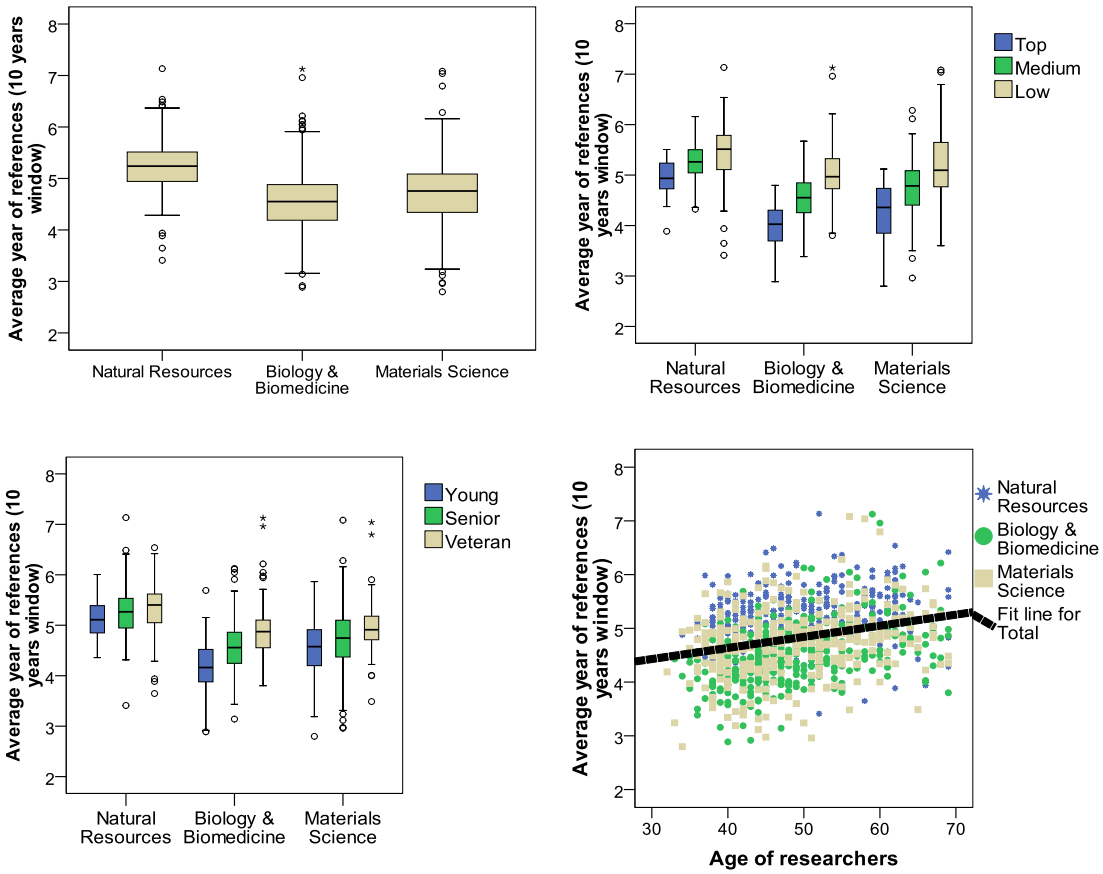


**Figure 5. Distribution of the average age of cited references by area.**

As observed before, Natural Resources researchers (followed by Materials Science researchers) present the oldest references, while Biology & Biomedicine presents the youngest ones ($p<0.05$ in all cases). By scientific performance classes and age clusters the same pattern as before is detected (statistical significant differences have been found in almost all cases, $p<0.05$; with the exception of senior and veteran researchers in Natural Resources). The graph on the bottom-right side of Figure 5 also supports the idea that older researchers tend to use older literature. Thus, we can claim that veteran and Low class scientists tend to use older literature as compared to younger and Top class researchers. This suggests that veteran researchers may have a better knowledge of older literature because that was the literature that was published when they were younger.

## Discussion and Conclusions

This paper is framed within the study of the possibilities of bibliometric indicators beyond the research evaluation purposes. In this sense, we have analyzed different aspects related with the use of information (i.e. cited references) by individual researchers from the perspective that references are key elements in the communication of scientific results and new ideas.

Our results show that the number of references per document is field-related. This finding is consistent with previous studies, although most of them analyse the issue at the meso level (see for example Albarrán & Ruiz-Castillo, 2011). A more interesting result is the fact that within each area, differences in the use of scientific literature according to the scientific performance and age of scientists are observed.

*Characteristics of Top researchers regarding the use of references in their publications*

In our analysis a clear observation is made: top researchers use in their papers a broader range of scientific literature as compared to the other researchers. They cite more references per document -also more references in their single-authored publications-, use more external references (i.e. references that do not belong to any of the co-authors of the papers) and they also use a larger list of unique references in their total production. These results can support the affirmations by Rames Babu & Singh (1998) that to be a productive scientist is important to be aware of what other scientists are doing, and that an acquaintance with recent trends of research is inevitable for raising one's own research output. These results have also implications from the perspective of library and information access policies, as they must provide tools in order to facilitate the access to the new knowledge published in the research fields of scientists, and thus allowing them to catch up with the last developments and ongoing research in their scientific areas (Ramesh Babu & Singh, 1998).

*Use of non-WoS material by researchers and age of the cited literature*

From a general perspective, differences among areas in the use of non-WoS literature and in the age of the cited material have been found. In this sense researchers in Natural Resources tend to cite a greater share of publications that are not covered by the Web of Science as well as older literature (similar results in natural resources-related fields have been also stated by Velho & Krige, 1984), which can be related with the local orientation of some of its research topics (Costas & Bordons, 2005). The case of Materials Science presenting shorter reference lists is in line with the claims of Kidd (1990) that engineering fields present less extensive (or comprehensive) bibliographies for their works. On the other hand, Biology & Biomedicine researchers present a more WoS-oriented trend in their referencing practices (this can be also linked to the fact that this is a very internationally oriented area) as well as a greater focus on more recent literature. These findings are in line with the results of Butler & Visser (2006) who showed how the coverage of the different fields through the Web of Science is different and that these differences can be inferred from the reference analysis of the publications from these fields.

From the individual perspective we can see how younger and Top researchers tend to use more WoS-covered literature, and discarding any population effect (here we have to remind that almost all researchers were already active in the first year of the study), this could be related to the different focus of the research topics of researchers, with the more veteran ones focusing on more local topics and using more books or other literature not covered by Web of Science.

In summary, it is clear that older researchers tend to rely on older literature, while the contrary holds for younger researchers. This was also an observation by Gringras et al (2008) who found that after the age 40, university professors in Quebec cited increasingly older literature, so older researchers were more distant from the forefront of research. Barnett & Flink (2008) suggested two potential explanations for this phenomenon: first, an age bias in the receptivity of scholars to new ideas, with younger scientists being more receptive than older ones; and second, the accumulated nature of knowledge explains that scientists start building their base knowledge when they begin their professional careers, so this base is older for older scientists.

Some interesting implications for research management and library policy can be derived from our research. On the one hand, according to all the previous statements, we can say that a good knowledge of the literature in the field facilitates scientists' performance and can help them to develop successful research, which in the long term may result in publications in the best journals and becoming highly cited. We could hypothesize that collaboration between young and old scientists can be especially appropriate in science provided that the former are closer to the forefront of science and the latter may contribute with their experience and

know-how. On the other hand, our study reveals the importance of library policies that facilitate the access to the new knowledge published in the research fields of scientists to allow them to catch up with the last developments and ongoing research in their scientific areas.

## References


Albarrán, P. ; Ruiz-Castillo, J. (2011). References Made and Citations Received by Scientific Articles. *Journal of the American Society for Information Science and Technology*, 62(1), 40-49.

Alimohammadi, D.; Sajjadi, M (2009). Correlation between references and citations. *Webology*, 6(2), a71.

Barnett, G.A. & Flink, E.L. (2008). Impact of the Internet and scholar age distribution on academic citation age. *Journal of the American Society for Information Science and Technology*, 59(4), 526-534.

Budd, J.M. & Magnuson, L. (2010). Higher education literature revisited: citation patterns examined. *Research in Higher Education*, 51, 294-304.

Butler, L. & Visser, M.S. (2006). Extending citations analysis to non-source items. *Scientometrics*, 66(2), 327-343.

Calero, C., Buter, R., Cabello Valdes, C. & Noyons, E. (2006). How to identify research groups using publication analysis: an example in the field of nanotechnology. *Scientometrics*, 66(2), 365-376.

Clarke, M.E. & Oppenheim, C. (2006). Citation behaviour of information science students II: Postgraduate students. *Education for Information*, 24, 1-30.

Clements, K.W. & Wang, P. (2003). Who cites what? *The Economic World*, 79(245), 229-244.

Corbyn, Z. (2010). An easy way to boost a paper's citations. *Nature,* 13 August. *DOI:10.1038/news.2010.406*.

Costas, R. & Bordons, M. (2005). Bibliometric indicators at the micro-level: some results in the area of natural resources at the Spanish CSIC. *Research Evaluation*, 14(2), 110-120.

Costas, R. & Bordons, M. (2006). Algorithms to solve the lack of normalization in author names in bibliometric studies. *Investigacion Bibliotecologica*, 21(42), 13-32.

Costas, R., van Leeuwen, T.N. & Bordons, M. (2010). A bibliometric classificatory approach for the study and assessment of research performance at the individual level: the effect of age on productivity and impact. *Journal of the American Society for Information Science and Technology*, 61(8), 1564-1581.

Gingras, Y., Lariviere, V., Macaluso, B. & Robitaille, J.-P. (2008). The effects of aging on researchers' publication and citation patterns. *Plos ONE*, 3(12), e4048.

Gonzalez-Bambrila, C. & Veloso, F.M. (2007). The determinants of research output and impact: a study of Mexican researchers. *Research Policy*, 36, 1035-1051.

Jiang, Y. (2008). Locating active actors in the scientific collaboration communities based on interaction topology analyses. *Scientometrics*, 74(3), 471-482.

Kidd, J.S. (1990). Measuring referencing practices. *Journal of the American Society for Information Science*, 41(3), 157-163.

Lariviere, V.; Archambault, E.; Gingras, Y. (2006). The place of serials in referencing practices: comparing natural sciences and engineering with social sciences and humanities. *Journal of the American Society for Information Science and Technology*, 57(8), 997-1004.

Levin, S.G. & Stephan, P.E. (1989). Age and research productivity of academic scientists. *Research in Higher Education*, 30(5), 531-549.

Long, R., Crawford, A., White, M. & Davis, K. (2009). Determinants of faculty research productivity in information systems: an empirical analysis of the impact of academic origin and academic affiliation. *Scientometrics*, 78(2), 231-260.

Moed, H.F. (2006). *Citation analysis in research evaluation*. The Netherlands: Springer.

Oppenheim, C. & Smith, R. (2001). Student citation practices in an Information Science Department. *Education for Information*, 19, 299-323

Ramesh Babu, A. & Singh, Y.P. (1998). Determinants of research productivity. *Scientometrics*, 43(3), 309-329.

Roth, W.-M. & Cole, M. (2010). The referencing practices of Mind, Culture and Activity: On Citing (Sighting?) and Being Cited (Sighted?)'. *Mind, Culture, and Activity*, 17 (2), 93-101.
Shanmugam, A. (2009). Citation practices amongst trainee teachers as reflected in their project papers. *Malaysian Journal of Library & Information Science*, 14(2), 1-16.
Velho, L. & Krige, J. (1984). Publication and citation practices of Brazilian agricultural scientists. *Social Studies of Science*, 14, 45-62
Wallace, M.L., Larivière, V. & Gingras, Y. (2009). Modeling a century of citation distributions. *Journal of Informetrics*, 3(4), 296-303.
Weingart, P. (2005). Impact of bibliometrics upon the science system: Inadvertent consequences? *Scientometrics*, 62(1), 117-131.
Zuccala, A. (2006). Modeling the invisible college. *Journal of the American Society for Information Science and Technology*, 5(2), 152-168.