Current Comments

Lifetime Citation Rates

Number 2 January 14, 1980

For about twenty years now I have written and lectured throughout the world on citation indexing.[1] Without doubt, one of the most often asked questions concerns the problem of predicting how often a given paper will be cited. Until recently, all I could do was tell people what we knew about the number of papers that achieved a given citation threshold.

Of course, as Derek Price has often pointed out, the more a paper is cited in its early history the higher is the chance it will be cited in the future.[2] Papers that are cited infrequently in the first several years after publication have little likelihood of being heavily cited afterward. There are, of course, notable exceptions to this. I have published in *Current Contents*® a series of most-cited papers for each year, the last series covering 1976 papers.[3,4] Almost invariably, these papers go on to become superstars. I am now also involved in identifying the so-called slow starters.

Several years ago, a group of professors at the University of Pennsylvania were motivated to quantify citation information more precisely. Nancy L. Geller, John S. de Cani, and Robert E. Davies decided to take an actuarial approach. We are all used to mortality tables so why not tables for a paper's lifetime citation rate (LCR)? While one may not believe that citation frequency says anything about the quality of the work in question, there is a practical side to the exercise of computing lifetime citation expectancies. Because of the *Science Citation Index*® *(SCI*®*)* many people are interested in them.

Although LCR projections would have widespread interest among scientists, the vagaries of the publication process led to the appearance of the paper by Geller, de Cani, and Davies in *Social Science Research*,[5] where it is unlikely to be read by the very audience for whom it was intended. That's one of the reasons I'm taking this opportunity to call it to your attention. Should you want complete methodological details, you can obtain reprints from Robert E. Davies, Department of Animal Biology, University of Pennsylvania/H-1, Philadelphia, PA 19104.

Based on the number of citations to papers already published, the LCR model projects the *approximate* number of citations a paper is *likely* to receive in the 40 years following its publication. LCR estimates for individual papers are then combined to project the likely overall citation rate for all papers a particular scientist has published to date. The authors believe that the LCR model measures the impact of a scientist's typical work over its lifetime. However, it should be stressed that the LCR model is an approximate, not exact, calculation.

The LCR model is based on what we know about general citation patterns and the annual growth of scientific literature. According to the authors, if you want to estimate the total number of citations a paper is likely to receive 40 years after publication, you must first determine the probability of its being cited in each of those 40 years. Using data recorded in the 1974 *SCI*, the authors present a graph of the chronological distribution of citations for the years 1966-74. The graph shows that citations follow a "quick growth and slower decay" pattern. For example, nearly 12% of all citations recorded in the 1974 *SCI* were to papers three years old. Nearly 6% were to papers seven years old and nearly 3% were to papers 11 years old.

Also, the authors point out that variations between different scientific fields affect the probability of a paper being cited some time after it was published. The annual growth rate of scientific literature varies for each field. For example, biochemical literature is now growing at an annual rate of 5.3%.[6] This means that the volume of biochemical literature will double in 13.4 years. In contrast, botany literature has an annual growth rate of 3%, or a doubling time of 25.5 years.[6] The faster moving field, biochemistry, relies less on the older literature than a field like botany. Table 1 shows estimated LCRs for papers in these two fields. Corrected for field size, the chance of a 20-year-old biochemistry paper being cited is less than that of a 20-year-old botany paper. This is not to say that biochemists don't cite older papers. In fact, their citation of older literature has increased in recent years.[6]

Taking all this into account, the authors generate a mathematical factor which adjusts a paper's citation probability both for years in the future and the past if the paper appeared before the *SCI* was first published in 1961. Also, variations between scientific fields are adjusted by different factor values for doubling times of eight, ten, fifteen, and "infinite" years and for each year from 1935 to 1974. I won't repeat here the details of the method—the authors provide the reader with thorough footnotes and exhaustive references.

The procedure used to figure out the lifetime citation rate of even a single paper is so complex that putting the LCR model to work is no easy task! The authors explain that for a complete citation history all of an individual's existing papers must first be compiled. This requires a lot of time and painstaking effort. To make things easier, the authors suggest starting with a complete bibliography before using the *SCI*. A complete bibliography should list all papers on which the individual scientist appears as primary or secondary author. Also, it helps you to avoid the homograph problem; authors with the same names who write on different topics can easily be differentiated.

After identifying all of an individual's papers, each one is examined for any irregular citation patterns: "A self-citation rate of more than about 10% of the number of references in the work is

Table 1: Estimated lifetime citation rates for papers in biochemistry and botany. Estimates predict the citation rates papers published in various years will have 40 years into the future. As you can see, papers in biochemistry will receive more citations than papers in botany although they started out with the same number of citations. This is due to the difference in "doubling times" for the two fields and the difference in average number of references per paper.

Year of Publication	1967	1965	1963	1961	1959
Number of Citations to Date	5	7	10	12	15
LCR Estimate for Botany Paper	7.9±2.2	9.7±2.0	12.7±1.8	14.7±1.8	23.5±3.7
LCR Estimate for Biochemistry Paper	10.2±3.3	12.1±3.0	15.4±2.9	17.3±2.8	25.6±4.2

unusual and should be noted. An investigation should also be made to see whether any of the citations negate or are critical of the cited work. The number of citations received by each paper in each year should be examined to find out if they fit the usual curve of quick growth and slower decay.... [Unusual citation patterns are also] associated with some successful books and with the very small percentage of papers describing techniques that become widely used."[5]

Assessing individual contributions to multi-authored papers is difficult. So *four* LCRs are actually calculated for multi-authored papers. "LCR all" credits the scientist for all citations to all papers, whether authored singly or jointly. "LCR per author" divides citations to each paper by the total number of authors on that paper. "LCR independent" credits the scientist for independent citations, i.e., citations to papers not coauthored by scientists of higher academic rank. "LCR independent per author" divides citations to independent papers by the total number of authors.

I warned you that it takes a lot of time and effort to calculate LCRs from complete citation histories! Jim Dolby, president of Dolby Associates in Los Altos, California, suggests that the LCR model needs to be simplified. "It would be nice to see someone make a much simpler estimate. Pick any sort of simple-minded model that would be a reasonable thing to consider for projection. Take someone's earliest three papers and average the number of citations over the first five years. Say that's his citation rate per paper and drop it right there. How good would the predictions be of some absurdly simple thing like that compared to the more complex [LCR] model? Presumably, the more complex model would do measurably better. So, if I were to follow up on the procedure I would go for a much larger sample [than the authors provide] and try two or three very elementary prediction procedures, and then show that this more elegant one worked better, if in fact that's the case."[7] In fact, the authors recently have simplified their model to avoid counting citations to *each* of a scientist's papers. Instead, an LCR approximation is calculated from the *total* number of citations to a scientist's work and the dates of each paper's publication.[8]

A number of other flaws must be ironed out before the LCR model can be confidently applied. John Tukey, professor of statistics at Princeton University, disagrees with the authors' formula for calculating the variance and standard error of LCR projections. The alternative formula Tukey uses gives results that differ from the authors' calculated standard error by as much as 50% in places. However, Tukey notes that this is a question of interpretation that doesn't have a right or wrong answer. "The authors might not agree with me...about the variance even after they've thought it over. I could see different answers with different interpretations of what you are trying to put a standard error on."[9]

Another problem in predicting the lifetime citation rate of a paper is how to account for the status of the journal in which it appears. Although it is possible to weight citation counts to reflect the prestige of a journal,[10] it is unclear how the weight should be used. John Tukey says, "If people publish in journals that get lots of citations [and thereby increase the chances of their papers being cited] we ought to downgrade them something for that. But you should also think these are probably the 'better' journals and you should probably upgrade them for that. I'm not at all sure whether we have the data to finish thinking [this problem] through, and which way we'd come up."[9] Geller, de Cani, and Davies recommend a separate assessment of journal status, but conclude only, "It might be useful to build this aspect of citations into further work on this topic."[5]

I'd like to point out that the LCR model predicts the future citation rate of *papers* but says nothing about the scientist's future creative potential. The LCR model is a means of predicting how often an *already recognized* paper will be cited. If we systematically gather enough data we may be able to predict, after several years, that a particular work is taking off—that is, it is an idea whose time is finally coming.

I'm excited by the comparative and predictive abilities of the LCR model, or whatever future shape the model will assume after its shortcomings are resolved. Science policy administrators would find such a model valuable once they know how to use it.

The obvious implication of such a model, if it were truly accurate, is that it enables one to compare the relative impact of more established scientists' work to that of younger researchers who haven't had the time to be recognized by their peers.

An LCR-type model might be a useful adjunct to the co-citation method of describing the structure of science, developed at ISI®.[11,12] In co-citation analysis, active areas are identified by counting the number of times two papers are cited together by later papers. If we can estimate the number of citations each co-cited paper is likely to amass over a 40-year period, maybe we can forecast the future development of now active specialties. Not only would we have a map of the scientific world as it appears today, but we would be able to speculate on how it might appear in the future.

* * * * *

My thanks to Linda Cooper and Alfred Welljams-Dorof for their help in the preparation of this essay.

REFERENCES

1. **Garfield E.** *Citation indexing—its theory and application in science, technology, and humanities.* New York: Wiley, 1979. 275 p.
2. **Price D J D.** Networks of scientific papers. *Science* 149:510-5, 1965.
3. **Garfield E.** The 1976 articles most-cited in 1976 and 1977. Part 1. Life sciences. *Current Contents* (13):5-23, 26 March 1979.
4. ------------. The 1976 articles most-cited in 1976 and 1977. Part 2. Physical sciences. *Current Contents* (17):5-16, 23 April 1979.
5. **Geller N L, de Cani J S & Davies R E.** Lifetime-citation rates to compare scientists' work. *Soc. Sci. Res.* 7:345-65, 1978.
6. **Garfield E.** Trends in biochemical literature. *Trends Biochem. Sci.* 4(12):N290-5, 1979.
7. **Dolby J.** Telephone communication. 22 August 1979.
8. **de Cani J S.** Telephone communication. 4 September 1979.
9. **Tukey J.** Telephone communication. 1 August 1979.
10. **Pinski G & Narin F.** Citation influence for journal aggregates of scientific publications: theory, with application to the literature of physics. *Inform. Process. Manage.* 12:297-312, 1976.
11. **Garfield E.** ISI is studying the structure of science through co-citation analysis. *Current Contents* (7):5-10, 13 February 1974.*
12. **Small H.** Co-citation in the scientific literature: a new measure of the relationship between two documents. *J. Amer. Soc. Inform. Sci.* 24:265-9, 1973.*

*Reprinted in: **Garfield E.** *Essays of an information scientist.* Philadelphia: ISI Press, 1977. 2 vols.