Citation Light Curve for the Very Large Array

S. R. Kulkarni

V1: 22 April 2014; revised March 2, 2015, primarily editorial comments.

1 Background

Of late I have been interested in the "business" side of astronomy. This interest results from the fact that the cost of astronomical facilities is increasing much faster than astronomer's ability to draw funds from public and/or private sources. In this environment it is only natural that stresses are developing in our community. A single large project can (and has) the ability to suck most of the available funding and thus starve mid-size science and PI-driven projects.

Occasionally it helps to have some data as inputs for decision making (I am being sarcastic). Of course, one could argue that scientific research is a matter of taste and thus not subject to bean counting methodology. However, big science (capital intensive) comes with its own baggage and metrics are inevitably a part of that. Furthermore, big projects or big surveys displace smaller projects and smaller investigations and so it is quite reasonable to consider the consequences.

These issues concern me in my capacity as the Director of Caltech Optical Observatories. With a fixed budget the dilemma is (1) do we build an "flagship" (necessarily expensive) instrument or (2) build cost-effective workhorses or (3) upgrade existing instruments or (4) improve the performance of the telescope (infra-structure). One could argue that organizations which are both stably funded and also plushly funded (e.g. ESO, ALMA, NRAO, Gemini and so on) also face the same problem but the problems I face are more acute and I have to take decisions far more decisively and on shorter timescales than large Observatories. [On the other hand, I can take the decisions after deep thought and consultation of a few key people and not be subject to many levels of scrutiny and be subject to pressures from multiple interest groups].

Motivated thus I have analyzed the scientific output of the Keck Observatory (on the eve of the 20th anniversary) and measured the productivity and impact of the different instruments of the Keck Observatory. I have some idea of the useful lifetimes of optical instruments and determined the quantitative benefit from upgrading existing instruments. From an optical perspective the VLA is a simple instrument (essentially a fixed backend). Having set up the MATLAB machinery for analysis I decided to apply it to the VLA – whence this informal report.

2 The Very Large Array

The Very Large Array is the world's premier decimetric array. It is located on the desolate Plains of St. Augustine in the state of New Mexico. The facility was formally inaugurated in 1980. The formal construction cost is quoted[1] to be $78.5. Assuming a mean epoch of 1973 the cost of the VLA, updated only CPI, is $78.5 × 5.29=$415M (FY2014). Despite the cost the VLA was and remains one of the productive and unique workhorses of modern astronomy. Here I investigate the astronomical returns of this venerable facility.

[1] Wikepedia

3 The NRAO Bibliometric Database & Interface

The NRAO librarian(s) have done an admirable job in building up a data base of NRAO papers.[2] I am informed that librarians pore through papers in journals and use a uniform criterion for including papers in the NRAO data base. The classification is quite detailed (key projects, archival research, papers arising from surveys etc).

Unfortunately, the interface to this wonderful data base has many limitations (in other worse, quite poor). The choices range from spartan (refereed or not; a given decade or the whole duration) to an extensive 70-strong list of Instruments (which I call as "collections"). There is no explanation for these instruments. It appears to range from the supreme (data arising from VLA observations) to minor projects (and traversing major surveys and substantial key projects). Some of these classifications are obvious but some ["Observational" with a non-trivial trove of 1312 papers] are not. A one-line explanation of these classes would be very very helpful.

I would like to clarify that the detailed classification is *extremely* useful for sophisticated analysis. The simplest use of the data base should allow the user to select Boolean operators – which is not the case. To compound this frustration the interface does not allow more than 100 papers to be displayed. As a result, to download the "VLA" data one would have to manually execute an inquiry 57 times. The reason for this parsimony in the displayed output is unclear.

Despite this irritants I decided, for the larger cause of radio astronomy, to soldier on and undertake an analysis of the "VLA" trove of papers. I downloaded[3] the displayed html output (as text files) for the following *Collections*: "VLA" (5665 papers), "eVLA" (268), "FIRST" (266), "NVSS" (442) and "Archival VLA" (608). The total is of these data sets is 7249 papers[4]. The rather larger number of html output files forced me to learn `bash` scripting. Each html file was filtered out to yield the `bibcode` and I built a file of `bib codes` for each *collection*. These `bibcode` lists saved as inputs to a MATLAB code which in turn queried the ADS engine and derived the `bibcodes` of the papers citing the VLA papers. The net result is that each VLA paper has an associated file consisting of the `bibcode` followed by the publication years of the papers citing that particular paper. Another MATLAB program then reads these files and does all the analysis reported below.[5]

4 Analysis

There are two measures for an Observatory: a "flux of papers" (the number of papers published in year "t"), $\mathcal{N}(t)$ and a "flux of citations" (the number of citations that the Observatory papers received in year t), $\mathcal{C}(t)$. The traditional and commonly used citation measure is to add up the number of citations attributed to papers published in year t, $c(t)$. This is an ill-defined measure since re-evaluating $c(t)$ a year from now will result in a different $c(t)$. For further discussion of the virtues of $\mathcal{C}(t)$ and pitfalls of $c(t)$ the reader is referred to Kulkarni (2014).

5 Results: VLA

The light curve of the citations is presented in Figure 1. In the Appendix I present a table of all the three measures. It is noticeable that the citation curve is rising more rapidly than linear. The daily flux of VLA, stands at about $44/\eta$ citations where η is the fraction of time that the array is available for astronomical observations.

[2] `https://find.nrao.edu/papers/`

[3] circa mid April 2014

[4] I verified that there are no overlapping papers between the data sets.

[5] The MATLAB program produces the tables, figures and even some text for direct incorporation into the .tex file. Over the years I have found this approach, though time consuming in the beginning, is a highly efficient way to write papers.

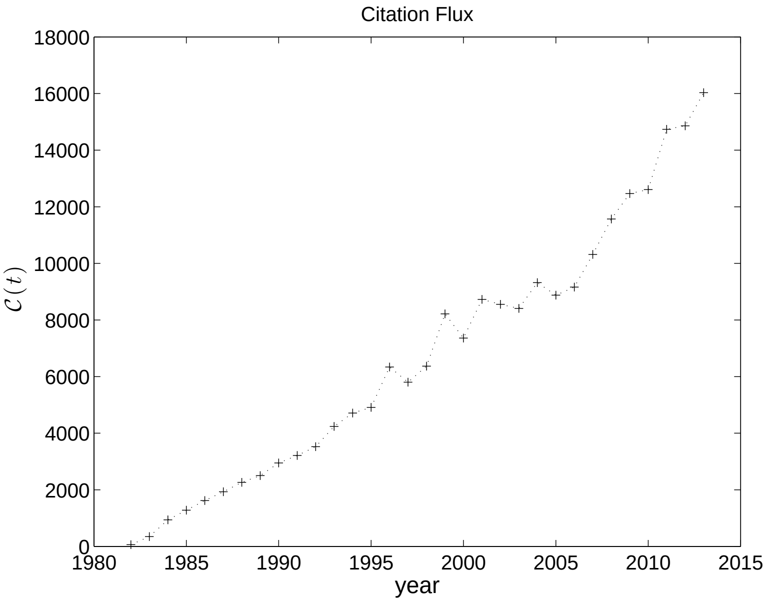

Figure 1: The citation flux arising from VLA refereed papers. The collections include "VLA", "eVLA", "ArchVLA", "FIRST" and "NVSS".

The paper flux and citation flux curve are arguably a good measure of the productivity of bread-n-butter science. However, these two measures are unlikely to capture the singular achievements of an Observatory. To this end, it is useful to look at the most cited papers. I have determined that papers with more than 600 citations stand out sufficiently above the tail of the citation histogram (Figure 2). Therefore I classify VLA papers with citations, n_c greater than 600 as "highly cited" papers. There are 10 such papers. These are ranked with the highest cited paper (NVSS - J. Condon et al) assigned Rank, R=1 (see Table 1).

The highly cited papers (in descending rank) are:

1. The NRAO VLA Sky Survey
2. The FIRST Survey: Faint Images of the Radio Sky at Twenty Centimeters
3. Evidence for a black hole from high rotation velocities in a sub-parsec region of NGC4258
4. A superluminal source in the Galaxy
5. A Redshift Survey of the Submillimeter

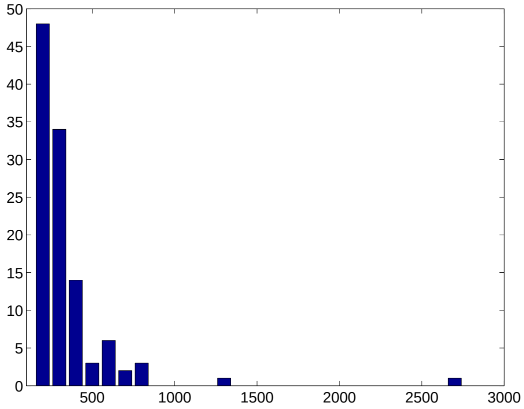

Figure 2: Histogram of the number of citations of highly cited papers. From a visual appearance I choose to call papers with citations, $n_c > 600$, as "highly cited" papers.

R	bibcode	n_c	year	coll.
1	1998AJ....115.1693C	2664	1998	VLA
2	1995ApJ...450..559B	1300	1995	VLA
3	1995Natur.373..127M	832	1995	VLA
4	1994Natur.371...46M	820	1994	VLA
5	2005ApJ...622..772C	755	2005	Archive
6	1991ApJS...76..813S	740	1991	VLA
7	1989ApJS...69..831W	692	1989	VLA
8	1989AJ.....98.1195K	621	1989	VLA
9	1992Natur.355..145W	613	1992	VLA
10	1990ApJS...72..567G	603	1990	VLA

Table 1: Highly cited papers.

Galaxy Population
6. The Einstein Observatory Extended Medium-Sensitivity Survey. II - The optical identifications
7. The morphologies and physical properties of ultracompact H II regions
8. VLA observations of objects in the Palomar Bright Quasar Survey
9. A planetary system around the millisecond pulsar PSR1257 + 12
10. The Einstein Observatory Extended Medium-Sensitivity Survey. I - X-ray data and analysis

5.1 All Sky Surveys

Given the renewed interest in sky surveys I plot the citation light curve of FIRST and NVSS in Figure 3. Apparently the citation light curves for such major surveys rise linearly for the first decade and then rapidly thereafter and reach a peak a decade later. The two surveys are the two most highly cited (NVSS, $n_c = 2664$ and FIRST, $n_c = 1300$). It is worth noting that that the two surveys are contributing a handsome 20% of the daily flux of citation. NVSS was granted 2700 hours (Condon et al. 1998). The total number of citations from NVSS now stands at 8249. Including the citations to the NVSS survey paper the "returns" from NVSS is 4 citations per hour. The total VLA citations now stand at 228,949. Assuming a 30-year duration with $\eta = 0.7$ I find the mean citation production for the VLA is 1.24 per hour. NVSS is thus three times more efficient in generating citations relative to the usual mix of observations.

Given that the surveys are now peaking it is not unreasonable for NRAO to consider a new survey. The time allocated for the surveys should be balanced against a daily flux of 44 new citations.

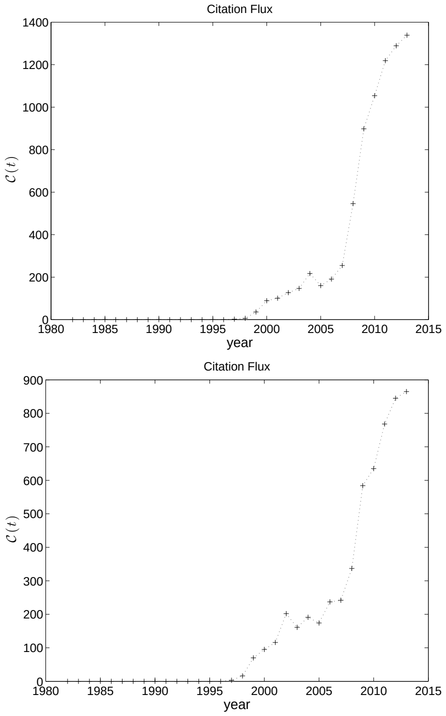

Figure 3: The citation flux arising from papers attributed as "NVSS" (top) and "FIRST" (bottom). The total number of citations for NVSS is 8249 and that for FIRST is 5867.

6 Should the VLASS being undertaken?

My answer is YES. The success of NVSS and FIRST shows the value of large area surveys. The empirical data is simply overwhelming. A survey which goes a factor of 10 deeper is now due. Time for such a survey can be easily obtained by placing a moratorium on the endless specialized deep looks for logN-logS astronomy (e.g. deeper and deeper looks at a dozen places around the sky such as CDFS, CDFN, HDF, Groth Strip, the Las Vegas Strip and so on). Such a reference image will be most valuable for all on-going and future optical Time Domain (TD) synoptic surveys. Similar all sky surveys are planned in the Southern hemisphere (ASKAP and perhaps MeerKAT). VLASS can be the trail blazer.

A personal note: I had the distinct pleasure of being the sole night-time observer for the VLA for a period of six months (June-August of 1981, followed by a few additional months). I had developed a phase-array technique (Kulkarni 1981, VLA Observer's handbook) which did not require the DEC-10 for computing (but only the Modcomp!). The result was that I was the beneficiary of more than 1200 hours of array time. I lived in a trailer on the site and the highlight of the week was Thursday dinner in Datil (vegetarians welcome for that day). J. van Gorkom, a newly minted PhD, had just arrived and thus began our life-long friendship. The resulting thesis paper was published as Dickey, Kulkarni, van Gorkom & Heiles (1983). This paper has accrued 73 citations or 0.04 citations per hour – which is most definitely the lowest citation returns per hour of array time!

7 Appendix

7.1 Suggestion for collections

I suggest the following top level classification (for the NRAO Librarain):

1. VLA (Direct use of data).
2. NVSS
3. FIRST
4. Arch-VLA Using data from VLA archive but excluding NVSS, FIRST)
5. eVLA (Direct use of data).
6. Arch-eVLA (using archival data from eVLA).
7. KeyProjects: COSMOS, THINGS, MASIV, LITTLE-THINGS, BIG-THINGS ...

7.2 Table of the annual fluxes

The data which went into Figure 1 is based on Table 2.

Year	$\mathcal{N}_p$	$\mathcal{C}(t)$	$c(t)$
1982	78	60	3804
1983	129	347	5390
1984	136	937	6603
1985	162	1278	6520
1986	158	1619	6010
1987	171	1931	6852
1988	167	2263	5505
1989	183	2503	8794
1990	174	2951	7961
1991	158	3215	7745
1992	184	3522	7724
1993	193	4241	7868
1994	206	4713	8625
1995	217	4910	9561
1996	192	6340	6761
1997	181	5802	7392
1998	197	6367	9207
1999	251	8216	7918
2000	212	7362	7853
2001	231	8727	8430
2002	238	8551	8097
2003	209	8407	6554
2004	253	9320	6792
2005	235	8876	7979
2006	251	9162	6946
2007	327	10316	9450
2008	359	11569	9556
2009	336	12466	5594
2010	298	12605	4855
2011	352	14736	4929
2012	282	14858	2631
2013	266	16032	1053

Table 2: Paper flux (N_p), citation flux ($\mathcal{C}$) & $c(t)$.

Note: Kulkarni (2014), has, of the date noted above, not been submitted.