Evaluating High Impact Papers: Are We Missing Something?

Sherry Winkelman[1],*, *Arnold* Rots[1], and *Raffaele* D'Abrusco[1]

[1]*Harvard-Smithsonian Center for Astrophysics*

Abstract. The bibliographic science papers with high citations rates are often used as an indication of the science impact of an observatory. These high impact papers are presented as examples of the best science being done with an observatory's data. But, is the number of citations by itself a good indicator of the scientific impact of the paper, and is impact a good indicator of the scientific impact of the observatory? In this paper we will present results from a recent study of Chandra high impact papers and suggest some alternative methods for identifying such papers.

1 Introduction

Observatory science papers with high citations rates are often used by observatories as an indication of the science impact of an observatory. Classically, the papers with the highest number of citations are deemed High Impact Papers (HIPs). In the case of the Chandra X-ray Observatory (CXO) we typically list the 50-100 papers with the highest citation counts (the top 1-2% of papers) and no further evaluation is done.

One reason for identifying HIPs is that High Impact Papers may tell us something about the most influential types of science coming from the observatory and therefore what sorts of observing programs will have long-lasting contributions to astronomy. Since it often takes many years for papers to reach the HIP category, we hope to identify objective measures which will allow us to identify High Impact Papers earlier to perhaps provide guidance in the request for proposals process.

2 What Makes a Paper HIP?

There are two components for defining a High Impact Paper: 1) evaluation of the science contribution from the observatory to the paper as a whole and 2) objective measures based on citation history and for which HIPs are outliers for the observatory. The first component is subjective, but the idea is that the science content from the observatory needs to be integral to the findings in the paper. Some reasons we have found for rejecting a paper as being a HIP are:

---

*e-mail: swinkelman@cfa.harvard.edu ORCID: 0000-0001-7354-6221

- the only connection to Chandra was to create a figure to confirm measurement of data from another observatory and the other observatory data was the focus of the paper

- cross-matching of sources in Chandra data with sources from another observatory but no further analysis using Chandra data is done

- X-ray data for only a few targets of a large sample of targets came from Chandra

- Chandra provided data showed a non-detection

- the Chandra analysis is in the supplementary materials

The objective measures are based on citation histories of papers. There are two caveats associated with the objective measures: 1) they may change as the observatory ages, so one must decide what to do about HIPs that no longer meet the current criteria and 2) they need to be determined in a programmatic way without requiring further (or minimal) analysis by curators.
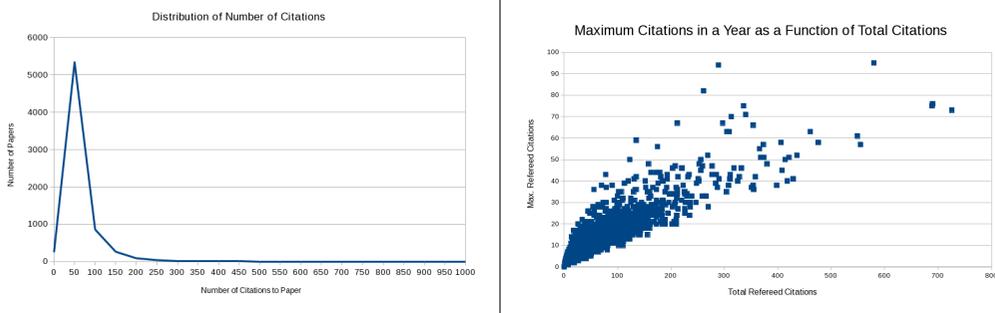
3 Morphology of HIPs

Figure 1. The plot on the left shows the distribution of number of citations in Chandra Science Papers. The long tail of the distribution starts at ∼ 200 citations. The plot on the right shows the distribution of the total number of refereed citations to the maximum number of refereed citations in a single year. More than 95% of CSPs are clustered in the lower left of the figure with total citations < 200 (corresponding to the long tail papers in the left plot) and maximum annual citation rates < 30, suggesting that papers with a maximum annual citation rate > 30 may be high impact.

We looked at four objective criteria for identifying potential High Impact Papers: 1) the distribution of the number of citations in a paper: indicates papers with high citations compared to the full body of science papers from the observatory (shown in left panel of Figure 1); 2) the number of citations in a paper versus the highest citation in a year: highlights papers with unusually high citations of a paper in a single year (shown in right panel of Figure 1); 3) the number of papers with > X citations in a year: provides papers which have a moderate citation rate for several successive years (shown in left panel of Figure 2); and 4) the number of papers which previously published results from a particular paper: gives you papers whose results are frequently used for subsequent studies (shown in right panel of Figure 2).

Based on the four criteria which we examined, we chose four categories for identifying potential High Impact Papers and identified a fifth category which shows promise as a category but the data linking in the Chandra Bibliography is not complete enough to set any limits.
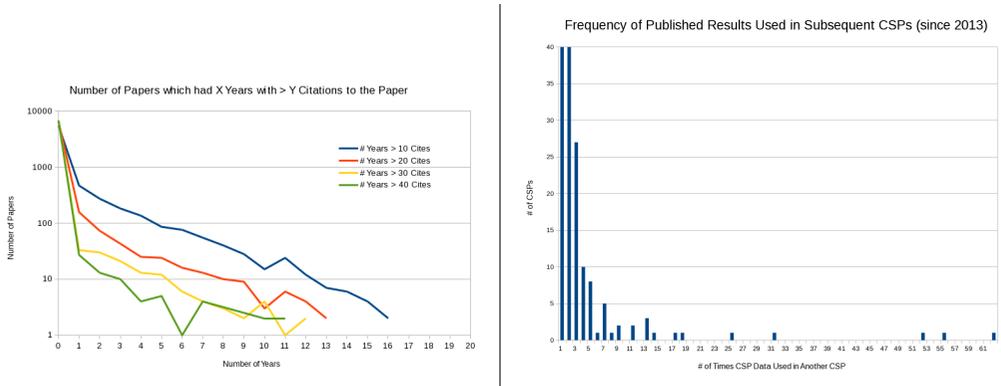
Figure 2. The left panel shows the number of papers which had X years with > *Y* citations to the paper. 5 and 10 years (∼ 1/4 and ∼ 1/2 the age of the Chandra archive) were chosen to as benchmark times. Categories C and D were somewhat arbitrarily chosen as the points were the curves fall below 10 papers to encompass those papers which steady growth rates in citations. The right panel shows the frequency that published results of the HIP are used in subsequent Chandra Science Papers. The data linking required for this plot are complete in the Chandra Bibliography for papers published after 2013.

- **Category A**: > 200 refereed citations is based on the left panel of Figure 1. The long tail of the distribution starts at ∼ 200 citations.
- **Category B**: > 30 citations in a single year is based on the right panel of Figure 1. Papers with a peak citation > 30 is in the scatter portion of the distribution.
- **Category C**: > 10 years with > 10 citations in a year is based on the left panel of Figure 2 and corresponds to steady citation rate for ∼ 1/2 of the mission.
- **Category D**: > 5 years with > 20 citations in a year is based on the left panel of Figure 2 and corresponds to a moderate citation rate for ∼ 1/4 of the mission.
- **Category E**: > *X* Chandra Science Papers (CSPs) have used the results from the HIP for the analysis in the CSP is based on the right panel of Figure 2.

The left panel of Figure 3 shows the distribution of the various flavors of HIPs for Chandra Science Papers. Also shown in the figure are the numbers of papers in each category which fulfill the science relation criterion. While most HIPs fit into more than one category, each category identifies some papers that do not fall in another category. Note that Category E is not included in Figure 3. Preliminary results suggest that this is a promising category, but the indirect data linking required for this analysis is not complete in the Chandra Bibliography. Reference [1] gives a description of the new data linking being added to the Chandra Bibliography.

The hope is that the new categories of High Impact Papers will help us to identify HIPs earlier. This is clearly seen in the right panel of Figure 3. Category B in particular can be used to identify HIPs as early as 2-3 years after publication, while HIPs in Category C can be identified as early as 5-7 years after publication.

4 HIPs Compared to Full Bibliography

Other than citation histories, are there other differences between High Impact Papers and the complete bibliography of Chandra Science Papers? We looked at four characteristics of HIPs versus CSPs:
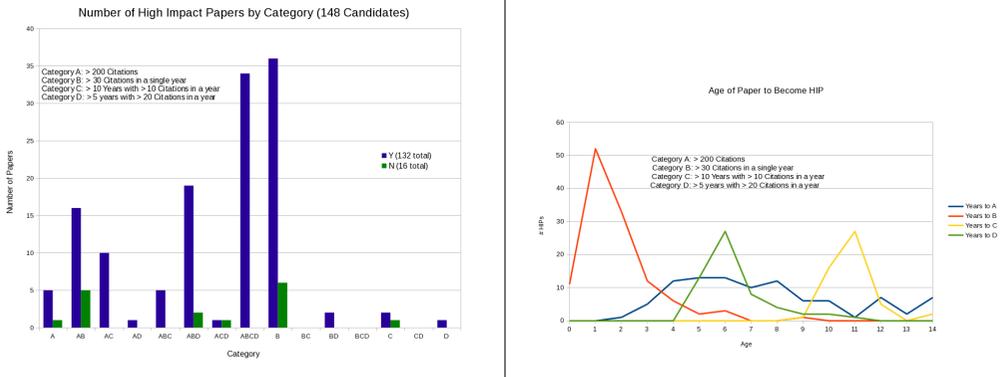
Figure 3. The left panel shows the number of HIPs by category combination. The blue bars indicate the number of HIPs where the Chandra analysis in the paper is integral to the findings in the paper. The green bars are potential HIPs for which the Chandra analysis is not significant to the paper. See Section 2 for examples of what falls within the green bars. The right panel is a plot of the age of a paper when it became HIP for a given category.

1. percentage of non-CSP citations in the paper

2. primary data usage in the paper

3. data from other wavebands also analyzed in paper
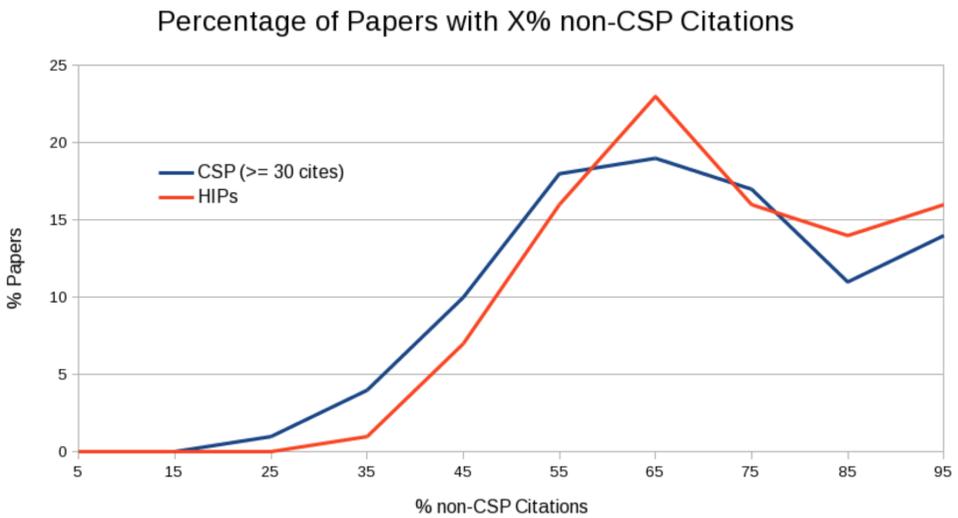
4. Chandra survey regions covered in paper

Figure 4. Plot of the percentage of papers with X% non-CSP citations. The blue line represents CSPs with >= 30 citations and the red line represents HIPs.

Data Comparison	CSPs	HIPs
Direct data analysis	79.5%	69.7%
Indirect data analysis	20.5%	30.3%
Multi-observatory	57.9%	52.8%
Theory/computation	10.4%	13.4%
Follow-up analysis	6.8%	6.3%

Table 1. Comparison of the primary Chandra data usage between CSPs published since 2013 and HIPs. Indirect data analysis is broken down into multi-observatory, theory/computation, and follow-up analysis.

Observatory Usage			Waveband Usage		
Observatory	HIPs	CSPs	Waveband	HIPs	CSPs
XMM	16%	22%	Visible	75%	30%
HST	13%	11%	X-ray	62%	46%
Spitzer	10%	5%	IR	23%	15%
VLA	9%	6%	Radio	19%	13%
VLT	7%	2%	UV	7%	5%
Swift	7%	8%	Submm	2%	1%
ROSAT	5%	2%	Gamma	1%	2%
Subaru	4%	1%	Microwave	1%	0%

Table 2. Other observatory data analyzed in CSPs published since 2013 and HIPs.

Figure 4 indicates that HIPs tend to be cited by non-CSPs more frequently than CSPs which could indicate that science in High Impact Papers is influencing non-Chandra science more than general CSPs.

Table 1 indicates that HIPs are more likely to be based on indirect analysis of Chandra data than CSPs as a whole. We flag three types of indirect analyses: multi-observatory analysis means previous Chandra results are used in a complementary fashion with data from another observatory; theory/computation means previous Chandra results are compared to models or used as input to theoretical computations; and follow-up analysis means previous Chandra results are as the starting point for further data analysis. For both CSPs and HIPs, multi-observatory analysis is the primary form of indirect analysis of Chandra data.

The Chandra bibliography also flags whether data from other observatories was analyzed in the paper. We directly flag a number of observatories and if an observatory is not in our list of observatories, we flag the waveband of the data used in the paper. We see in Table 2 that there are no significant differences between HIPs and CSPs which also have observatory flags when looking at the top cited observatories, but we do see a significant difference when looking at wavebands. HIPs combine analysis of Chandra data with data from visible and other X-ray observatories at a much high rate than CSPs as a whole.

Table 3 shows that HIPs are far more likely to be connected with a deep Chandra survey or catalog than CSPs which are connected to a survey. A secondary reason that these surveys are frequently used maybe the homogeneity of the sample. Not surprisingly, those surveys are also those areas of the sky which are covered by deep surveys in other wavebands.

Survey	HIPs	CSPs
Chandra Deep Field South	50%	24%
Chandra Deep Field North	43%	8%
C-COSMOS	12%	19%
Extended CDFS	7%	6%
Chandra Large Area Synoptic Survey	5%	0%
Chandra Source Catalog	3%	4%
Chandra Orion Ultradeep Project	3%	2%
AEGIS-X	3%	6%
X-Bootes	3%	2%

Table 3. Survey regions covered in CSPs published since 2013 and HIPs.

5 Requirements for Inclusion in Bibliography

To make HIPs an integral part of the Chandra bibliography, there will need to be additional fields in the database and code development for identifying potential HIPs. In the database, at minimum we will need to add a flag that the paper is considered to be high impact. To be more useful, we need to flag when a paper is a candidate for being a HIP, which HIP categories the paper qualifies for, and when the paper qualified for that category. Another useful piece of information to have associated with a HIP is a brief description about the paper to include with replies to requests for lists of HIPs.

Procedurally, when a potential HIP comes in, a curator will need to determine whether the paper should be considered a HIP or not based on an evaluation of the science contribution of Chandra data to the paper as a whole. The HIP flag should then be set to the appropriate value and additional metadata added to fully classify the HIP.

It should be noted that potential Category E HIPs will also require a determination by curators of whether the referencing papers are part of a series of papers from the authors rather than independent work. This may be a significant burden to the system and requires careful consideration.

Finally, one needs to deal with HIPs which no longer fit the current criteria. Some considerations are: once a HIP, always a HIP; a HIP with qualifications; or perhaps, HIP versions.

6 What Next?

During the course of this study we identified three areas which need deeper investigations before moving forward on adding HIPs to the Chandra bibliography:

- Perform a historical analysis of objective criteria at ages 5, 10, and 15 years of the Chandra mission to see how they may have changed and whether 'former' HIPs are still HIPs

- Determine if the age of median citations as a function of the age of the paper help inform the HIP determination

- Explore predictive traits based on citation history which could flag potential HIPs

7 Summary

In summary, we have found that there are two components in defining a High Impact Paper, a subjective evaluation of the science contribution from Chandra to the paper as a whole and an objective

determination based on citation history. Based on citation distributions of CSPs we identified four categories for identifying potential HIPs and that two of those categories allow us to identify High Impact Papers much earlier than using the total count of citations.

When comparing HIPs to the full set of Chandra Science Papers, we find that: HIPs tend to be cited by non-CSP's more frequently than CSPs, perhaps because HIPs tend to use data from other wavebands more frequently than CSPs; HIPs are more likely to be based on indirect analysis of Chandra data than CSPs; and that HIPs are far more likely to be connected with a deep and/or homogeneous Chandra survey than CSPs connected with surveys.

Inclusion of HIPs into the Chandra bibliography will require expansion of the metadata connected with paper and code development. A deeper understanding of how the objective criterion change over the course of the mission is needed before determining the extent to which HIPs should be included in the bibliography.

References

[1] S. Winkelman, A. Rots, R. D'Abrusco, *Growing a Bibliography*, These Proceedings (2018)